Running Head: Measurement and Item Response Theory

Measuring Constructs in Family Science:
How Can Item Response Theory Improve Precision and Validity?

RACHEL A. GORDON *University of Illinois at Chicago*

Published in: *Journal of Marriage and Family, 77*, 147-176 (February 2015).

Department of Sociology and Institute of Government and Public Affairs, University of Illinois at Chicago, 815 West Van Buren St., Suite 525, Chicago, IL 60607 (ragordon@uic.edu).

Key Words: measurement, methods, theory.

I gratefully acknowledge funding from the Institute of Education Sciences (R305A090065, R305A130118) and the National Institutes of Health (R01HD060711) as well as very helpful comments from Ken Fujimoto and Everett Smith. All errors and omissions are my own.

Abstract

This article provides family scientists with an understanding of contemporary measurement perspectives and the ways in which item response theory (IRT) can be used to develop measures with desired evidence of precision and validity for research uses. The article offers a nontechnical introduction to some key features of IRT, including its orientation toward locating items along an underlying dimension and toward estimating precision of measurement for persons with different levels of that same construct. It also offers a didactic example of how the approach can be used to refine conceptualization and operationalization of constructs in the family sciences, using data from the National Longitudinal Survey of Youth 1979 (n = 2,732). Three basic models are considered: (a) the Rasch and (b) two-parameter logistic models for dichotomous items and (c) the Rating Scale Model for multicategory items. Throughout, the author highlights the potential for researchers to elevate measurement to a level on par with theorizing and testing about relationships among constructs.

Constructs are fundamental ingredients of family theories, forming the building blocks of research questions and hypotheses (White & Klein, 2008). An essential component of quantitative research is the operationalization of such concepts, many of which are difficult to observe. As a consequence, the reliability and validity of instruments are central concerns of family researchers. Although most family scientists would likely agree with these statements, reviews continue to periodically lament problems with the definition and operationalization of key constructs in the field. In this article I consider the ways in which item response theory (IRT) can help scholars determine whether they have precisely and validly assessed constructs of interest. A focus on IRT is important given that many scholars are trained primarily in classical test theory (CTT) approaches. As I discuss, IRT is not a silver bullet: It cannot solve all measurement challenges in the field, and in many cases CTT can be leveraged to meet similar goals. IRT does, however, feature some aspects of measurement that are less apparent in the ways that applied scholars often use traditional CTT approaches and, as such, has the potential to help scholars rethink how they approach the task of defining good measures.

I begin by providing a rationale for this review, highlighting key recent publications. I then discuss overarching principles of measurement that encompass IRT and CTT. Doing so is meant to broadly frame IRT within contemporary perspectives about sound measurement. I emphasize three contemporary orientations: (a) unified validity, (b) conceptual frames, and (c) local validation. I then offer a nontechnical introduction to some key features of IRT and offer references to direct readers who want to learn more. Next, I provide an empirical example, demonstrating some of the ways that IRT can offer new insights. I aim to highlight the potential for researchers to elevate measurement to a level on par with theorizing about and testing of relationships among constructs. I end by encouraging publications of theoretical and empirical

research on measures in mainstream family sciences journals.

Why Cover This Topic?

Before getting into specifics of measurement principles and IRT approaches, it is helpful to lay out a rationale for the importance of this topic. As already noted, most family scientists would likely agree that good quantitative science requires reliable and valid measures. Indeed, most Method sections in journals like the *Journal of Marriage and Family* (*JMF*) include some information to make the case that the measures are good assessments of their underlying constructs, and authors often point back to scale developers' publications, reporting internal consistency estimates for the sample in hand.

A challenge to the field, however, is that scale development and validation are not as integrated into the published literature as are studies that relate scale scores to one another with regression models. When unpublished, an instrument's development and refinement do not benefit from the critical peer-review component of the scientific process. Even when published, measurement articles more often appear in specialized methods journals than mainstream, substantively oriented journals, making them less visible to family scientists and somewhat divorced from theory development. As a consequence, measurement risks becoming a side exercise rather than a central and substantial component of a research project. Likely contributing to the separation of measurement from mainstream science is the fact that psychometric theory has advanced rapidly in recent decades, with new models that can seem quite different from (and more technical than) familiar CTT approaches. Indeed, the limited training on measurement built into many disciplines' graduate programs exacerbates the challenge to scientists of staying abreast of this advancing field (Aiken, West, & Millsap, 2008; Aiken, West, Sechrest, & Reno, 1990)

Perhaps it is not surprising, then, that Blinkhorn (1997) lamented that psychometric tools such as factor analysis are often mere data-reduction techniques and internal consistency reliability estimates tend to be used in a perfunctory way. Writings in other fields suggest this issue is not unique to particular disciplines. Several decades ago, Schwab (1980) noted that theoretical advances in organizational studies were hampered "because investigators have not accorded measurement the same deference as substantive theory (and) as a consequence, substantive conclusions have been generated that may not be warranted" (p. 34). In the field of criminology, Piquero, Macintosh, and Hickman (2002, p. 521) concluded that "Researchers have become too complacent and uncritical about the measurement of their key dependent variable, [self-reported delinquency]." And, in a review of articles published in top criminology journals between 2007 and 2008, Sweeten (2012) identified only five of 130 studies that used IRT approaches.

Reviewing articles published since 2000, I likewise identified a limited but growing set of exemplary studies, including in prominent family science journals (e.g., Bridges et al., 2012; Browning, 2002; Browning & Burrington, 2006; Fincham & Rogge, 2010; Funk & Rogge, 2007; Krishnakumar, Buehler & Barber, 2004). Several studies have demonstrated the potential ways in which IRT can be used to create shortened versions of scales that are nearly as informative as longer versions (Cole, Rabin, Smith, & Kaufman, 2004; DeWalt et al., 2013; Piquero et al., 2002); another showed how IRT suggested fewer response categories were needed than in an original scale (Osgood, McMorris, & Potenza, 2002). Reducing the number of items and categories in these ways, while maintaining precision of measurement, can reduce cost and response burden in studies. Several studies also highlighted the extent to which items were well targeted at the populations of interest. In many cases, items were concentrated in one pole of the

construct (with little or much of the construct), with fewer items at the other pole (Osgood et al., 2002; Piquero et al., 2002). IRT models helped show how estimates based on these measures were imprecise for people in the regions of the construct lacking items, leading to reduced power for detecting associations. Tests of differential item functioning also showed that in some cases items operated differently across groups, sometimes leading to consequential changes in substantive conclusions; such studies often examined problem behaviors and delinquency (e.g., Cho, Martin, Conger & Widaman, 2010; Conrad, Riley, Conrad, Chan, & Dennis, 2010; Gibson, Ward, Wright, Beaver, & Delisi, 2010; Piquero et al., 2002; Rocque, Posick, & Zimmerman, 2013; Raudenbush, Johnson, & Sampson, 2003) but also family processes (Bingenheimer, Raudenbush, Leventhal, & Brooks-Gunn, 2005; Krishnakumar et al., 2004).

Even with this emerging literature, the number of measurement studies that I identified was dwarfed by regression-based articles, and measurement articles continued to be more often found in methodological than in substantive journals. In the remainder of this article, therefore, I aim to introduce IRT methods to a broader range of family scholars in an effort to help the field fully harness the potential of the latest measurement perspectives and psychometric tools.

Selected Principles of Measurement

The importance of sound measurement has long been emphasized in the physical and social sciences. Particularly important to family scholars is the considerable development throughout the 20th century and into the 21st century of new ideas and tools for validating measures. Many of these developments were relatively rapid, and they did not always diffuse widely beyond methodologists and certain substantive subfields. As a consequence, family research has not benefited as much as it might from new conceptualizations and approaches.

Three professional associations encompassing the fields of psychology, education, and

measurement published standards for testing and measurement throughout the latter half of the 20th century that served as major compilations of advances in the field and helped organize and codify emerging conventions. The 1999 standards were used for more than 15 years (Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education [hereafter *Joint Committee on Standards*], 1999) and a revision was published as this article went to press (Joint Committee on Standards, 2014). As recently as 1974, the standards articulated several types of validity that remain commonly used in practice and publications, such as construct validity, content validity, and predictive validity. The more recent editions offered a revised, unified conceptualization of validity. Under this perspective, validity does not come in different discrete types. Instead, evidence accumulates about different aspects of validity, and the full complement of evidence should be weighed when making decisions about particular uses of the measure. In short, this contemporary perspective eschews blanket statements that a measure is "valid" or "not valid" (or that it "does" or "does not" demonstrate, say, "predictive validity"). It is important to note that particular kinds of evidence may be more important for certain uses of a scale, and the totality of evidence may be weighed somewhat differently for one use versus another.

A measure of depressive symptoms, for example, would require particular precision (small error bands) around a specific cutoff level if it were to be used for clinical referrals. On the other hand, if the measure were to be used as a continuous outcome in a large nationally representative study, then adequate precision of measurement across the full range of the scale would be desirable. Conclusions like these may seen intuitive in hindsight, but scholars do not always lay them out explicitly when making choices about which measures to use in a study. Too

often, a scale is used in multiple ways based on one set of evidence that claims the scale is uniformly "reliable and valid." As I describe below, an important benefit of IRT is quantifying precision of measurement not only for a scale as a whole but also in particular ranges of the instrument (e.g., around a clinical cutoff), which can help scholars better evaluate the evidence for different uses. In short, the Joint Committee's standards defined validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" and "it is the interpretations of test scores for the proposed uses that are evaluated, not the test itself" (Joint Committee on Standards, 2014, p. 11).

Several additional points emphasized by the Joint Committee's standards suggest ways in which family scientists can leverage IRT models to improve measurement. One of these is the need to fully articulate a conceptual framework for a measure. As already noted, the need for such conceptualization is often recognized by family scientists, but recurring complaints about fuzzy constructs and conceptual overlap suggest that the field is not focusing enough attention on this aspect of measure development. IRT offers family scientists new ways to think about the extent to which a set of items represents a construct and to iteratively refine and revise their conceptualizations, especially in regard to how well the items cover the full range of the latent construct (of course, factor analysis is another important tool for test refinement, and I describe below some of the relationships between the two approaches). Papers that fully articulate and theoretically ground such conceptualizations about measures, and then rigorously examine them empirically, have the potential to elevate to the level of other conceptually grounded work that empirically examines regression relationships among measures. When they do so, such measurement papers may more often be publishable in top substantive journals.

The Joint Committee's standards also emphasize the importance of local evidence—

evidence that an instrument is valid for the context in which it is (or will be) used—either obtained through existing data gathered from similar contexts or from new data gathered in the new context. Together with a movement away from blanket statements that a measure "is" or "is not" valid, the emphasis on local validation encourages researchers to replicate evidence about various aspects of validity across studies. In the same way that meta-analyses increasingly accumulate evidence about regression coefficients, scholars can systematically synthesize evidence about various aspects of validity (e.g., R. A. Peterson, 2000; Shafer, 2006). Not only can such syntheses document the extent of validity evidence across local contexts, but also this new evidence can importantly complement evidence created by a scale's developers (who may have financial and professional stakes in a scale). A shift in mindset toward replication and local validation may encourage editors and reviewers to be more responsive to articles that examine various aspects of validity in new samples or that systematically synthesize such evidence.

Some Strengths and Limits of IRT

This article does not offer the space or the context for a complete introduction to IRT (for more detailed and comprehensive treatments, see Bond & Fox, 2007; de Alaya, 2009; Embretson & Reise, 2000; Reise, Ainsworth, & Haviland, 2005; and Wilson, 2005). I therefore offer a fairly nontechnical introduction to a few distinguishing features of the approach, in particular in contrast to CTT approaches. I highlight the ways IRT can help scholars (a) scrutinize the positioning of items along the underlying latent construct and consider how their relative ordering aligns with theory and (b) examine precision of measurement across the range of the latent trait (rather than assuming a single precision level applies to the full range of the underlying scale). As I discuss below, CTT approaches can provide some of the same kinds of information, although typically they are not as clearly revealed in standard usage.

It is useful to start by describing overarching orientations toward measurement. From traditional perspectives, family scientists sometimes view items as largely interchangeable replications that signal a true score. Different items within a single test or parallel forms of a test are "direct evidence, each with the same weight and the same scope of coverage, about the same true score" (Mislevy, 1996, p. 385). From this vantage point, the internal consistency of a set of items or high correlation among parallel forms of a test are of utmost importance given that "A high reliability coefficient indicates that a different sample of tasks of the same kind would order the examinees similarly, which would lead to the same decisions about most of them" (Mislevy, 1996, p. 386). With this orientation, scholars may tend to write highly similar items and have little guidance about how to differentiate among them. It is also possible that researchers may too easily write items with content that overlaps other constructs, unless construct definitions are quite precise and differentiated. The reliability coefficient will not alert the researcher to these problems, to the extent that the largely redundant items or overlapping constructs are correlated.

Although researchers can use classical approaches to help define constructs and select items, theoretical and empirical construct maps used in some IRT approaches can shift how scholars approach measurement development (Wilson, 2005; Wright & Stone, 1979). Traditional factor-analytic models also assume continuous items, and IRT models are better suited to the Likert-type scales common in family science, modeling how the probability of a particular score on a particular item depends on an underlying latent continuum. Dichotomous items offer the simplest starting point, although the model extends to items with multiple ordered or unordered categories (e.g., below I demonstrate one approach, the Rating Scale Model, and discuss the strengths and limits of this and other approaches). Although pedagogical examples often focus on individual behavior, IRT models can be applied to measures of characteristics at various

levels, including individuals' attitudes and processes in couples or organizations. For instance, Gordon, Fujimoto, Kaestner, Korenman, and Abner (2013) examined a Likert-type rating scale of classroom quality in child care centers, including teacher–child interactions.

To begin with a simple case, consider a measure of violence that includes dichotomous items responded to with a "yes" or a "no" for behaviors such as "has thrown rocks, bottles, or other objects at another person" or "has attacked someone with the intent of hurting or killing them" in the last 12 months. In the IRT framework, the probability of a "yes" on such an item depends on the positioning of both the person and the item on the underlying violence construct. A youth who is positioned higher on the construct (is more prone to violence) will be more likely to respond "yes" to each item than a youth who is positioned lower on the construct (is less prone to violence). Our objective in measurement typically is estimating these positions of the youth on the construct (known in academic achievement testing as *person abilities*). The basic IRT framework also holds that all youth will be more likely to respond "yes" to an item that is positioned lower rather than higher on the construct (e.g., throwing rocks or bottles vs. attacking to hurt or kill). In academic testing and many other applications it is natural to refer to lower positioned items as "easier" and higher positioned items as "harder"; thus, item positioning is often described as the item's *difficulty level*. Other terms include *severity level*, or simply *item location*. I generally use *item difficulty*, given conventions for labeling the relevant item parameter in psychometric models and related software, but I sometimes use the more neutral *item location* when interpreting my empirical example.

More precisely, the simplest IRT model for dichotomous models—the Rasch model—defines the probability of an affirmative response to an item as a function of the difference between the position of the person and location of the item on the underlying dimension, with the

functional form of the model being the logistic distribution familiar to many readers; that is:

$$P(X_{si} = 1|\theta_s, \beta_i) = \frac{\exp(\theta_s - \beta_i)}{1+\exp(\theta_s - \beta_i)}, \quad (1)$$

where X designates an item, θ_s is the position of person s on the underlying dimension, and β_i is the location of item i on the underlying dimension (Embretson & Reise, 2000, p. 67). The fact that the basic model is a logistic function has several important implications, including that associations between response probabilities and the underlying construct are nonlinear and that it is natural to embed the Rasch model within a multilevel logistic regression model (which is being done increasingly; e.g., Raudenbush et al., 2003). Under Equation 1, a person has a 50% chance of responding affirmatively to an item that is positioned at her ability level. As the positive difference between the person's position and the item's position increases—she is positioned increasingly higher on the latent trait than the item such that the item is relatively "easier" for her—she is more likely to respond affirmatively. If she is positioned below an item, then she will have less than a 50% chance of responding affirmatively (the item will be relatively "hard" for her). Later in this article I provide figures that illustrate these associations.

With this orientation in mind, scholars can approach the writing and evaluation of items differently than is often done in the family sciences. In particular, under the IRT framework items are no longer fully interchangeable with one another. Instead, items are thought of as falling at different positions along the underlying continuum, much like marks fall at different intervals along a ruler. As a consequence of trying to place the items along such a ruler, scholars are pushed to think hard about the definition of a construct and how items operationalize the construct. The IRT model offers feedback with empirical estimates of the items' positions on that ruler. Such feedback can be used to refine the conceptual framework and its operationalization.

Although some analysts consider item difficulties like these from a CTT perspective, IRT

models estimate the location of items and persons (or other units, e.g., couples or organizations) on the same scale, allowing their relative positioning to be revealed. All else equal, an item whose difficulty is positioned at the same level as the person will be most informative for estimating that person's position on the underlying construct (Embretson & Reise, 2000, p. 184). Items that are very easy (positioned well below) the person or very hard (positioned well above) the person would be least informative. For representative population studies, the IRT orientation suggests that items would typically be desired that are well dispersed across the full range of the underlying dimension. This would ensure that items exist that are near the position of most people in the population (and therefore near the position of people in the sample drawn from that population). Gaps along the dimension that lack items would be undesirable because there would be less information for estimating the position of people in that range. On the other hand, if a particular sample focuses on one range of the underlying population—a sample of violent youth in my example—then a scale with items concentrated in that range of the dimension would be desirable. Although once articulated these statements seem fairly obvious, the IRT orientation sharpens attention to them and, importantly, the Rasch model (and other IRT approaches) provides estimates of the precision with which a scale estimates locations of people along the underlying dimension (e.g., a scale designed specifically for violent youth would be estimated to offer little information to distinguish among youth with little tendencies toward violence). Below I demonstrate how scholars can use IRT approaches to gain such insights.

The Rasch model is a *one parameter logistic*, or *1PL* model, referring to its one item parameter (the difficulty level, β_i). A common alternative is a *two-parameter logistic*, or *2PL* model. This model adds a second item parameter, known as discrimination, and is written as follows (with α_i representing the new term):

$$P(X_{si} = 1|\theta_s, \beta_i, \alpha_i) = \frac{\exp[\alpha_i(\theta_s - \beta_i)]}{1 + \exp[\alpha_i(\theta_s - \beta_i)]}. \quad (2)$$

Whereas the item difficulty parameter determines location on the underlying construct—analogous to an intercept—the discrimination parameter is analogous to a slope. In fact, the underlying logistic curve associating increases in a person's position on the underlying dimension with his or her increasing probability of responding "yes" to an item will be steeper for an item with higher discrimination. In other words, two people positioned at two different levels of the underlying dimension will be more likely to respond differently in the steepest region of an item with high discrimination than of an item with low discrimination. Because of the *S* shape of the underlying logistic function, however, items with high discrimination—high slopes—will offer these distinctions among people in a narrower range of the underlying dimension than would be true of less discriminating items (i.e., the item with high discrimination will offer relatively little information for people positioned farther from the item's difficulty level).

There are two important points to keep in mind about the discrimination parameter. One is that the 1PL (and Rasch) model assumes that all items are equally discriminating. The equal-discrimination assumption simplifies interpretation. If items differ in discrimination then there will not be a single order of items on the underlying dimension. In other words, the underlying logistic *S* curves will cross and, in some ranges of the underlying construct, Item 1 will be harder than Item 2, whereas in other ranges Item 2 will be harder than Item 1. This complicates conceptual interpretation (e.g., what would we make of a finding that youth at some levels of violence are unlikely to "throw rocks or bottles" but are likely to "attack people to hurt or kill"?). It is helpful for applied scholars to understand—but not be distracted by—a debate among psychometricians between the 1PL/2PL and Rasch models. Some characterize the debate as a

contrast between, on the one hand, finding the model that best fits the data and, on the other hand, designing a test that conforms with model assumptions (e.g., Andrich, 2004). Regardless of an analyst's perspective on this issue, testing the equal-discrimination assumption can be informative in the development and refinement of an instrument. I illustrate both the Rasch (1PL) and 2PL models below, featuring both software and conventions from the Rasch tradition and general purpose software and test statistics.

A second important point related to the discrimination parameter is that it is analogous to a factor loading in factor analysis, where it is often interpreted as the item's importance to measuring the underlying dimension (Reise, Widaman, & Pugh, 1993). When the mean structure is modeled in factor analyses, the item difficulty is parallel to the item intercept (Brown, 2006). Indeed, the lines between factor analyses and IRT models are less distinct than often thought (Takane & de Leeuw, 1987; Wirth & Edwards, 2007), and the interrelationships between the two approaches are being increasingly recognized as unidimensional IRT models have been extended to multidimensional IRT models (Gibbons et al., 2007; Muraki & Carlson, 1995) and as factor-analytic models have been extended to dichotomous and polytomous outcomes (Bock, Gibbons, & Muraki, 1988; B. Muthén, 1983; Skrondal & Rabe-Hesketh, 2007). Still, scholars trained in standard approaches to using factor analysis may emphasize factor loadings, focusing on retaining those items with factor loadings above standard cutoffs in refining instruments; to analysts used to this standard factor-analytic perspective IRT models can offer novel ways of thinking about the ordering of items on the latent construct.

Before I turn to an example, it is helpful to keep in mind some things that IRT does not do and some important model assumptions. One important point is that the Rasch model does not produce summary scores (estimates of persons' positions on the underlying dimension) that are

markedly different from simple raw sums of the same items; that is, scores produced from Rasch models are highly correlated with raw sums of the same items, and associations with other variables are highly similar (e.g., see Osgood et al., 2002, and Sweeten, 2012, for examples). Researchers often find this result surprising, but it is less so if one understands that the raw sum is a sufficient statistic for the underlying latent trait for the Rasch model. Indeed, strong proponents of the Rasch model see this result as a strength: The model assumes that if items indicate a single underlying dimension (is unidimensional) and contain no extraneous information (are locally independent) then the raw sum provides all that is needed to estimate the person's most likely response pattern (Embretson & Reise, 2000). For example, if the Rasch model holds, then a person who scores a 3 on a 10-item test would be most likely to have responded affirmatively to the three easiest items on the test and responded negatively to the seven harder items.

Given the high correlations between Rasch scores and raw sums, researchers may question whether Rasch is worth the trouble. One response to this is that Rasch scores and raw sums are similar but not identical, and seemingly small differences between them can matter. In general, Rasch models are nonlinear transformations of the items and, as such, Rasch scores and raw sums will differ most at the extremes. These differences have been shown to be consequential in some circumstances, and even small differences can offer a better estimate of true associations (and thus more statistical power). For example, Fraley, Waller, and Brennan (2000) showed the advantages of IRT scaling when measuring adult attachment, and other studies have found that the importance of using raw scores versus IRT estimates depended on item and person characteristics (Culpepper, 2013) and could be more consequential for more complex models (e.g., estimating interactions; Embretson, 1986; Kang & Waller, 2005).

More important, to the extent that scholars see IRT as a tool for measure development and refinement, then differences in results that are and are not informed by the approach would be more substantial; that is, when researchers scrutinize their construct definition and item content in preparation for a Rasch analysis they can gain clarity in their conceptual framework. Empirical results from a Rasch analysis can then be used to verify whether conceptual expectations are met or one should further refine the framework. Items that do not fit model assumptions, or that do not contribute much information, can be identified and replaced, or they can be refined to fit better or to offer more information. Scores from the resulting new items would then be expected to be only moderately correlated with scores from the original items, and the old and new scale scores would be expected to associate differently with covariates. In other words, although the Rasch scores and raw sums of the new scale would be expected to be highly correlated with each other (as with the Rasch scores and raw sums based on the original scale) both types of new scale scores would have been improved by information gleaned in the Rasch analysis.

Illustration of IRT Models

Scholars have devoted considerable attention to measuring children's behavior problems (Achenbach, 1991; Achenbach & Rescorla, 2001; J. L. Peterson & Zill, 1986) and adolescent delinquency (Elliott, Huizinga, & Ageton, 1985; Thornberry & Krohn, 2000). Both constructs have been widely examined in research on families, including through linkages to family structure (Fomby & Bosick, 2013; Osborne & McLanahan, 2007), parental employment (Coley, Ribar, & Votruba-Drzal, 2011; Johnson, Li, Kendall, Strazdins, & Jacoby, 2013), parents' mental health (Buehler, 2006; Turney, 2011), and parenting practices (Barnes, Hoffman, Welte, Farrell, & Dintcheff, 2006; McLoyd & Smith, 2002), among others. Measures of problem behaviors

have been widely analyzed with traditional factor-analytic techniques, providing consistent evidence of unique dimensions, at least at broad domain levels (e.g., externalizing vs. internalizing problems; Chorpita et al., 2010; Lambert et al., 2003). IRT analyses are more recent, but my review identified this as a subfield of rapid growth, perhaps in part because the potential array of the items along the latent dimensions of behavior problems and delinquency is fairly intuitive (e.g., Conrad et al., 2012; Lambert et al., 2003; Osgood et al., 2002; Piquero et al., 2002; Rapport, LaFond, & Sivo, 2009; Studts & van Zyl, 2013).

For the illustration, I relied on reports from mothers of the original National Longitudinal Survey of Youth 1979 (NLSY79; www.bls.gov/nls/nlsy79.htm) about the problem behaviors of their children in 2002. I chose these data because they offered a large public-use data set and readers could readily replicate and extend my analyses. The example data file, statistical programs, and annotated output are available as supporting information on the *JMF* website (http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1741-3737). Of course, the results I chose to present from this example were meant to illustrate some key features of IRT approaches—including differences in the way that analysts following the Rasch and 2PL traditions might approach the analyses and interpret the results—and are not meant to reflect all of the analyses and decisions that would go into a project meant to draw substantive conclusions. It is also the case that there is no single standard for presentation, especially given the paucity of published articles that have used IRT in the substantive literature, the variety of models and software used in the methodological literature, and the tendency for proponents of the Rasch and the 2PL approaches to publish within their own scholarly communities. I chose to highlight results and formats that I identified most consistently in my literature review and that I anticipated would be most helpful to substantive scholars. Readers interested in conducting their own analyses are

encouraged to consult the reference materials cited in this article and to examine published IRT studies in journals related to their subfield for additional guidance. Given the space constraints, I also could only scratch the surface of analytic tools for examining the fit of models and interpreting their results. Again, the references I cite offer many additional tools for interested readers to explore.

I first presented a set of results based on the Rasch model, including analyses of dichotomous items with the standard Rasch model and of multicategory items with the Rating Scale Model. I conducted these analyses with Winsteps, a user-friendly software program with a graphical user interface written to estimate a broad spectrum of Rasch-based models (Linacre, 2012). The software is available at a reasonable cost (less than $150, www.winsteps.com). A fully functional student version is available but is limited to 25 items and 75 cases. A free DOS-based version is also available, allowing thousands of items and cases. The manual provides tips for easily creating a Winsteps control file and data set from many different software packages, including SAS, SPSS, and Stata. I relied on the user-written Stata command *raschcvt* to easily convert my data file for Winsteps (Wolfe, 2001). The command creates a program ("control file") so that a basic Winsteps model can be easily run with minimal modifications (notes are available in the supporting information for this article on the *JMF* website).

I also estimated the 1PL and 2PL models using Stata's *gsem* command, which was introduced in Stata 13 for generalized structural equation models (including with nonlinear link functions; StataCorp, 2013). I reported fit and test statistics common to all maximum-likelihood models that can be used to evaluate overall model fit and to test the assumption of equal discriminations. Similar models can be estimated in other software packages, such as Mplus (L. K. Muthén & Muthén, 2012). As these advanced modeling options become increasingly

available, researchers will be able to flexibly use IRT approaches within familiar software, connect IRT with familiar factor-analytic models, and embed the latent construct within a broader structural equation model. Demand for such approaches should increase their functionality and usability. For instance, I relied on user-written code for interpretation in Stata (Ho, 2014), but Mplus has already added several traditional IRT graphics to their built-in commands. Other software packages are also adding macros and built-in commands for implementing IRT models using various estimation approaches, and readers are encouraged to check their favorite software for the latest developments (e.g., in SAS, the experimental PROC IRT [SAS Institute, 2013], as well as macros for PROC GLIMMIX [Black & Butler, 2012] and PROC NLMIXED [Sheu, Chen, Su, & Wang, 2005]; in SPSS, extension commands based on R [IBM, 2014]; in R, various user-written codes, including Chalmers, 2012, and Rizopoulos, 2006).

Testing Model Assumptions

The *residual*—the difference between the observed and expected response—lies at the heart of testing assumptions and fit of IRT models, similar to many other statistical techniques. As is common for dichotomous logistic regression models, for the Rasch model the observed residual is the difference between the observed response, x_{si}, which is either zero or one, and a predicted probability, p_{si}, calculated by substituting the estimated person and item locations into Equation 1. These residuals could theoretically range from −1 to 1. The extreme values are approached when a person's predicted probability approaches 0, but he actually endorses the item, or when a person's predicted probability approaches 1, but he does not endorse the item. Residuals may reflect random errors, such as a person becoming distracted, tired, or bored, or systematic errors, which might include an item reflecting a secondary dimension that affects responses for some or

all people. For instance, if an item contains a word that is unfamiliar to some respondents, they may be less likely to endorse it than others who are positioned at the same place on the latent continuum but are familiar with the word. Analysts can draw considerable information from close examination of the residuals. I focus on a subset of tools, primarily summary statistics at the item level. Readers interested in conducting their own analyses are encouraged to consult cited sources (e.g., Bond & Fox, 2007; Linacre, 2012; Wilson, 2005; Wright & Stone, 1979) that provide accessible guides to gaining insights from more detailed results, including the extensive output available from Winsteps.

Dimensionality. The basic Rasch and 2PL models shown in Equations 1 and 2 assume that the analyzed items measure a single dimension (although these models have been extended to allow multidimensionality, as I discuss below). Of course, defining single dimensions and writing items to capture them is challenging, and subfield experts sometimes disagree about dimensions. For instance, some scholars might consider behavior problems to be a single dimension; others might separate it into broad internalizing and externalizing domains; and still others into narrower subdomains such as anxiety, depression, aggression, and disruptiveness. Indeed, many problem behavior scales allow for multiple scorings at these different levels (e.g., Achenbach 1991; Achenbach & Rescorla, 2001).

I proceeded with a set of items from the Children of the NLSY79 that is often treated as a broad-band construct: the externalizing items from the Behavior Problems Index (BPI), which were completed by mothers of children ages 4 to 14½ in 2002. Item wording and descriptive statistics are provided in the Appendix. There were 16 items in total, and mothers were asked to report about the child's behavior over the last 3 months on a three-category response structure: (a) "not true," (b) "sometimes true," or (c) "often true." To illustrate the standard Rasch model

and its extension to the Rating Scale Model, I analyzed both the full three-category response structure as well as a dichotomous coding that collapsed together the sometimes true and often true categories. A sizable sample of 2,732 mothers completed the BPI, providing at least 18 cases in each response category (typically several hundred cases or more). The focal children were evenly split by gender (51% male) and were racially-ethnically diverse (20% Hispanic, 26% Black, 54% non-Hispanic, non-Black). The NLSY79 was based on a complex sampling design. I presented unweighted results for the illustration, but discuss below strategies researchers can use to adjust for weights and other design features.

Dimensionality assumptions are sometimes tested with factor analyses (appropriately specified for categorical items; e.g., Jöreskog & Moustaki, 2001) prior to estimating IRT models. Rasch analysts also commonly rely on item fit statistics, referred to as *infit* and *outfit*, as one way to summarize unexpected scores. The names refer to *in*liers and *out*liers, with infit more sensitive to unexpected responses by persons located near the item on the latent trait and outfit more sensitive to unexpected responses by persons located far from the item on the latent trait. Each is based on the sum of squared residuals and transformed to represent a mean square value (often labeled *MNSQ*) bounded by 0 and positive infinity. Values close to 1.0 reflect good fit.

Although psychometricians discourage strict reliance on cutoffs, MNSQ values between 0.70 and 1.30 are conventionally considered acceptable (Bond & Fox, 2007; Linacre, 2012). Standardized values are sometimes also reported; these values follow the normal distribution, and so results larger than 2 or 3 in magnitude are considered extreme. The unstandardized values may be preferred, however, when the number of cases is large because in these situations standardized values will be large even when infit/outfit values are not meaningfully big (A. B. Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). Different patterns of large or small values

on the infit and outfit statistics also suggest different problems with the items. One possible explanation for high values on both infit and outfit is the presence of a secondary dimension that affects all respondents; high values on just one of the measures may suggest that the secondary dimension is more salient to a subset of respondents (located near the item, in the case of infit, or far from the item, in the case of outfit; see Bond & Fox, 2007; Linacre, 2012; and R. M. Smith, 2004, for more guidance on diagnosing misfit). For instance, on the BPI, if the study investigators had inadvertently included an item that did not focus on behavior problems—perhaps how much the child enjoys playing sports—mothers' responses to that item would likely depart from model predictions.

Item fit statistics for the BPI dichotomous items are shown in Table 1. Given my large sample size, I focused on the unstandardized MNSQ values, which all fell within the suggested cutoff values of 0.70 and 1.30. In other situations, where fit statistics exceeded cutoffs, one might choose to restrict the analysis to a subset of items that best reflect a single construct (e.g., reanalyze a subset of items that further content review suggests may better reflect only the subdomains of aggression and delinquency). During instrument development and refinement, scale developers might also replace relatively poorly fitting items with other items, selected to better reflect the desired item content. One might also further scrutinize the residuals to identify the response patterns and other characteristics of persons who have large residuals on these items, to see whether the pattern of relatively larger residuals is unique to mothers or children who share particular characteristics (detailed Winsteps output facilitates such finely grained analysis).

<Table 1 about here>

Another way that Rasch analysts identify secondary dimensions is to factor analyze the

residuals (Linacre, 2012; R. M. Smith, 1996). If the items reflect a single dimension, then only random error should be left after accounting for the primary factor captured by the basic Rasch model. The presence of additional factors in the residuals would reflect lingering associations among items, beyond this primary factor. Some simulation studies suggest that values greater than 2.0 in eigenvalue units indicate a potentially meaningful secondary dimension (Linacre, 2012). Other simulation studies suggest other potential cutoffs (e.g., see Raîche, 2005), and analysts may also want to analyze the residuals in additional ways and consider additional simulations, as they evaluate dimensionality. In my example, the eigenvalues were at or below 1.6 for the BPI items, supporting my proceeding by treating the set of items as a broad externalizing dimension (of course, for a substantive article, I might choose to pursue additional residual analyses and item content review to consider subdomains, as discussed above).

Local independence. The assumption of local independence can be violated for various reasons, including when items share similarities in their stems, wording, or reference passages (e.g., if respondents read several paragraphs and then respond to a series of questions about each, then the responses for each set of passages will be correlated; Steinberg & Thissen, 1996; Wang, Cheng, & Wilson, 2005; Yen, 1993). Such dependencies can result in underestimates of standard errors and overestimates of item and test information (Wang et al., 2005; Yen, 1993). Especially when several items share a common feature, the dependency can be conceptualized as an additional dimension. I considered a commonly used summary statistic available in Winsteps: Yen's *Q3* statistic, which is designed to identify dependencies through pairwise correlations of item residuals. Common practice scrutinizes pairs of items with *Q3* values above .20 for potential dependencies, although simulation work is still accumulating regarding the appropriateness of this recommendation and the *Q3* statistic's performance relative to alternative

measures (Chen & Thissen, 1997; Ip, 2001; Houts & Edwards, 2013; see also citations below for more advanced models to identify and adjust for local dependence).

In my example, I identified one pair of BPI items with a *Q3* value above .20: the two items dealing with school: (a) "Is disobedient at school" and (b) "Has trouble getting along with teachers," with *Q3* = .26. The next highest *Q3* value, although below the suggested cutoff, identified the pair of items focused on other children: (a) "Has trouble getting along with other children" and (b) "Is not liked by other children," with *Q3* = .14. Given this slight evidence of local dependence, I retained all 16 items in the pedagogical example. In a substantive article based on these items, an analyst might choose to exclude these items or use one of the methods discussed below to adjust for local dependence. In iterative test development, these results might point scholars to consider the referent context of each item, given that the identified pairs deal with reports about the child's behavior at school and with peers. In my case, the NLSY79 investigators might have expanded the number of school- and peer-specific items, to form separate subscales, or focused only on general questions (with no referent) or explicitly on children's behavior at home.

Equal discrimination parameters. The S-shaped logistic function is a logical approximation for response probabilities on dichotomous or multicategory items. As with the use of logistic regression models for dichotomous outcomes, in the IRT context the logit link appropriately constrains predicted probabilities to fall between 0 and 1 (see, e.g., Gordon, 2012, pp. 564–565). In other words, the logit is the log of the odds, where the odds is the probability of a success (being scored a 1 on the item) relative to the probability of a failure (being scored a 0 on the item). Logit values are unbounded (can range between negative infinity and positive infinity), although in practice they typically fall between about −4 and 4. In the 2PL model the

characteristic *S* shape of the logistic function can take on a range of steepness across the items, from nearly horizontal to sharply increasing. In the 1PL and Rasch models the discriminations are constrained to be the same value across items (with the value being 1 in the Rasch parameterization).

The equality of discriminations can be tested formally through comparison of nested models with the *gsem* command, and the Stata results produce various fit values typically used with maximum-likelihood estimation (StataCorp, 2013). Because the underlying latent trait has an unknown scale, however, some constraints must be placed on model parameters for identification (this is consistent with other latent models, e.g., factor analysis; see de Ayala, 2009, for an accessible discussion). Analysts following the Rasch and 2PL traditions typically make different choices. In the standard Rasch model the scale is identified through the constraint that all discrimination parameters are 1. When presenting information to help consider the equal discriminations assumption, Winsteps approximates a constraint that the *average* of the discrimination parameters is 1 (with the *DISCRIM=YES* option; Linacre, 2012). In contrast, analysts following the 2PL tradition typically constrain the variance of the latent distribution to 1, rather than any of the discrimination parameters. The 1PL model can also be parameterized with the latent variance constrained to 1 and the discrimination parameters constrained to equal one another. Although equivalent parameterizations with constraints either on discrimination parameter(s) or on the latent variances will have the same model fit, the actual values of the discrimination (and difficulty) parameters differ and must be transformed (equated) to show their correspondence (de Ayala, 2009).

Results of some alternative specifications are presented in Table 2. I consider the parameter estimates below, and for now focus on the overall model fit, which is summarized in

the bottom three rows of the table. The three *gsem* models are in Column 2 (1PL), Columns 3 and 4 (2PL, with average discrimination of 1) and Columns 5 and 6 (with the variance of the latent distribution constrained to 1). As expected, the overall model fit of the alternative 2PL specifications was identical. I asked Stata to compare the 2PL and 1PL model with a likelihood ratio test, which was significant, $\chi^2(15) = 231.02$, $p < .001$, and the information criteria were also more than 10 points smaller for the 2PL than the 1PL model. Using Stata's postestimation commands, I then tested which pairs of discriminations differed significantly from each other. I found that for about 40% of pairs (50 of 120 comparisons), two discriminations did not differ from each other (at $p > .10$). With the latent variance constrained to one (Column 5 in Table 2), the discrimination parameters ranged from 1.18 to 2.26. Analysts from the 2PL tradition would take this as evidence that the 2PL model better captures the empirical range of discrimination parameters and would proceed with the 2PL model.

<Table 2 about here>

When I used the *constraint* option to constrain the discrimination parameters to average 1 in *gsem*, the discrimination parameters ranged from 0.65 to 1.25 (Column 3 of Table 2). As indicated with superscript letters in Column 3, six discriminations did not differ significantly from 1 in this parameterization. Analysts from the Rasch tradition might choose to focus on just these six items that best fit the model assumptions. Alternatively, as I illustrate below, analysts might conduct additional analyses to consider whether the discriminations are meaningfully different from 1 (especially given the relatively large sample in my example). Scholars might additionally scrutinize the other 10 items to understand how and why they differed from the rest, and perhaps revise the instrument to reflect relevant subscales (e.g., the five items with discrimination values less than 1 appear to encompass mood and restlessness, whereas the five

with values greater than 1 encompass irritability, cruelness, and temper, and those statistically equal to 1 disobedience, impulsiveness, and destructiveness).

Before proceeding, it is also instructive to note the equivalence of the alternative specifications of the Rasch/1PL and 2PL models. The difficulty estimates from Winsteps (Column 1) and *gsem* 1PL (Column 2) correlate at 1.00. The Stata results are centered on the persons rather than the items, however. The *gsem* item mean is 1.51, so the items are located 1.5 logits above the persons. The Winsteps person mean is −1.58, so the persons are located about 1.5 logits below the items. This reflects the fact that the absolute position is arbitrary (because the absolute zero point on the latent trait is unknown), although the relative position of persons to items is invariant. For the 2PL models, when parameterized so that the discriminations average 1, the variance of the latent distribution is estimated to be 3.25. When parameterized so that the latent variance is 1, the discriminations average 1.80, which is the square root of 3.25. As a consequence, the difficulties in the former parameterization are 1.80 times those in the latter (e.g., $-0.28 \times 1.80 = -0.50$ which matches, with rounding error, −0.51).

Interpreting Results

Rasch model. In this section, I exemplify two graphs commonly used in Rasch analyses: (a) the item characteristic curve and (b) the item–person map. The item characteristic curve (ICC) shows the probability that a person will endorse an item (in the dichotomous BPI items, that the mother will say a statement is "sometimes" or "often" true of her child), conditional on the latent trait (the child's latent tendency toward behavior problems). The item–person map displays the distribution of the persons (sampled Children of the NLSY79) and administered items along the latent continuum.

ICCs. Figure 1 shows an example of an ICC for the first BPI item, "Has sudden changes

in mood or feeling." The fitted curve has the expected *S* shape of the logit function. Its inflection point (where it turns from increasing at an increasing rate to increasing at a decreasing rate) occurs at its item location (difficulty), which the top left cell of Table 2 shows is −2.14 in the Winsteps parameterization. The conditional probability of 0.50 intersects the *S* line at this difficulty level. This means that a mother whose child is located at −2.14 on the latent dimension of problem behaviors would have a 50–50 chance of saying that sudden changes of mood or feeling are "sometimes" or "often" versus "not" true of her child. As the child's position on the latent trait increases, the mother's likelihood of endorsing the statement increases. For instance, by the logit value of 0 on the horizontal axis—approximately 2 logits above this item's location—the mother has a 90% chance of endorsing the item. As the child's position on the latent trait decreases, she is less likely to endorse the item. For instance, by the logit value of −4.0 on the horizontal axis—approximately 2 logits below this item's location—the mother has about a 10% chance of endorsing the item. These results visualize expectations based on Equation 1 regarding how the relative position of the person to the item affects the probability of a correct response in the Rasch model.

<Figure 1 about here>

Item–person map. Because the Rasch analysis constrains the items' discriminations to be equal, they all follow the same basic shape as the example shown in Figure 1. The way the ICCs differ is in their location; the curve shifts to the left or to the right, depending on each item's estimated location (difficulty). An instructive way to present these various item locations is in an *item–person map*. This map allows us to evaluate the empirical ordering of the items, from "easiest" (endorsed by many mothers) to the "hardest" (endorsed by few mothers). If I were able to predict, a priori, what the item ordering might be, based on theories or prior studies, then I

could evaluate whether the empirical ordering is consistent with my expectations. If conceptual and empirical work were insufficient to make such predictions, then I could use the map post hoc, to help refine my definition of the construct and my operationalization of it through these items and this instrument. It is important to note that the Rasch model (and other IRT models) places the persons on the same scale as the items. Thus, I am able to display the estimated locations of the children in my sample alongside the estimated locations of the items, helping to visually evaluate the extent to which the items are well targeted at the sample.

The item–person map from the Winsteps output is depicted in Figure 2. The persons appear on the left, and the items appear on the right. Winsteps arrayed the display so that the bottom represents children who had relatively less of the latent behavior problems construct and represents items that were relatively easier for mothers to endorse (were relatively frequent). The top part of the display reflects relatively more of the construct and relatively harder (or rarer) items. In the item distribution, the labels were abbreviated to accommodate the visual space of the graph, and the labels began with the item numbers for easy reference to the full labels in the Appendix. In the person distribution, each hash mark reflects 30 children, and each dot reflects one to 29 children.

<Figure 2 about here>

The scale on the far left side of Figure 2, labeled *MEASURE*, is in logit units. As noted above, in Winsteps the items are centered at a mean of 0, and therefore all item and person locations are relative to the item mean. This centering is reflected in the graph, and the letter *M* on the right side of the middle dashed line represents this item mean of 0. The letters *S* and *T* above and below this *M* represent 1 and 2 item standard deviations above and below the mean. Another letter *M* also appears on the left of the dashed line, below −1 on the logit scale. This was

the mean of the "persons" (the children in my sample; as shown in Table 2, the person mean is more precisely −1.58). The *S* and *T* on the left side of the dashed line represent 1 and 2 empirical standard deviations in the person distribution.

Analysts can draw meaning from each of the distributions on their own, as well as their distributions relative to one another. Because the items and persons were on a common scale, the map in Figure 2 makes clear whether the items were well targeted at the sample. In my case, the items were relatively "hard": They are concentrated in the upper end of the person distribution. One way to see this is by the fact that the person mean was more than 1 logit below the item mean. Another way to see this is by the fact that the person distribution appears skewed, with many children concentrated at the low end and a long tail at the upper end. Recall that each hash mark represents 30 children, so there are nearly 400 children located at a logit of −4 and more than 300 children located at a logit of −3, whereas just a handful of children are located at logit positions of +3 and +4. This result told me that relatively few children possessed some or much of the construct measured by these items; most children had very little of it. In the next section I illustrate how such item targeting affects item and test information.

Focusing on the item distribution, I also noted that the example item I show in Figure 1 is located at the bottom of the item map; its Winsteps item difficulty is −2.14. Its content—"sudden changes in mood or feelings"—was the most commonly demonstrated behavior problem among this set of items. "Arguing too much" was similarly placed on the latent construct. Nearly a logit higher was a cluster of three items—"stubbornness," "impulsiveness," and "disobeying at home"—followed by another set of three items: (a) "restlessness," (b) a "strong temper," and (c) "cheating or lying." Just above the item mean are the items capturing the child having "obsessive thoughts" and being "high strung." About a logit above the item mean is a set of items focused

on school and others: “bullying,” “disobeying at school,” and “not getting along with other children.” A bit higher are two more such items: (a) “not being liked by other children” and (b) “having trouble getting along with the teacher.” Finally, the hardest item was “breaking things deliberately.”

Careful scrutiny of this item map offered precise information for scale improvement and for conceptual insights. For instance, I saw that some areas of the logit scale had multiple items, whereas others had none. If items could be revised or replaced to move some from the clusters into the gaps, then the measure would be more informative. This would be helpful both for the internal gaps between clusters of items as well as the bottom of the scale where, although more than 700 children fell below −3 logits, no items occupied this range. I could also examine the item array for conceptual insight. Many orderings made sense—arguing and having mood swings were placed lower on the latent trait than having a strong temper; a strong temper, in turn, fell below breaking things deliberately, on the latent continuum. It is interesting that the items about school and other children were concentrated at the very high end of the logit scale. Earlier I showed how these may be locally dependent, and might be a separate subscale. The item–person map further suggests that scale developers might want to write “easier” items more reflective of relatively more common problems with behavior at school or consider whether mothers have sufficient knowledge and modest enough bias to accurately report about children’s behaviors with peers and at school.

Rating scale model. The results just presented are based on the dichotomous scoring that is sometimes used with the BPI items. I now turn to results for the full three-category response structure (see Appendix). I relied on the Rating Scale Model, a straightforward extension of Equation 1 for multicategory items. Whereas the Rasch model had a single item location

(difficulty) that represented the point on the latent trait where a person would have a 50–50 chance of endorsing or not endorsing an item, the Rating Scale Model estimates thresholds between each pair of adjacent categories where a person has a 50–50 chance of choosing the higher over the lower adjacent category. For the BPI items there were two adjacent pairs, and thus two thresholds, one reflecting the choice of "sometimes true" versus "not true" and the second reflecting the choice of "often true" versus "sometimes true." In the Rating Scale Model one assumes that the logit distances between the thresholds are the same across all items. The model estimates an overall location for each item, as well as two step parameters that are added to each overall item location to obtain the thresholds. Assumptions differ for other advanced models; for instance, some allow the thresholds and discriminations to differ across items, and some constrain the thresholds to be ordered (see de Ayala, 2009; Embretson & Reise, 2000; Linacre, 2012; Nering & Ostini, 2010).

In the following sections, I first present an example category probability curve from the Rating Scale Model, which shows the probability of each of the three responses conditional on the child's location on the latent trait and illustrates the locations of thresholds. I then present an item–person map similar to that depicted in Figure 2 but reflecting the broader coverage of the scale achieved by multiple categories. Finally, I show item and test information functions, which summarize the gains achieved by this fuller coverage of the latent trait.

Category probability curves. A category probability curve for the first BPI item is illustrated in Figure 3. In contrast to Figure 1, which shows a single curve to represent the conditional probabilities for the dichotomously scored item, Figure 3 has three curves representing the probability of the lowest category (*not true*, the solid black line that was downward sloping), the middle category (*sometimes true*, the gray line that peaked in the

middle), and the highest category (*often true*, the dashed black line that was upward sloping). At each point along the latent continuum the values of the three curves sum to 1. The highest curve at each point is the most likely category. Thus, in my example, mothers were most likely to select the option "not true" up until about −3 logits, where they became most likely to select "somewhat true," up until about 0 logits, when they became most likely to select "often true." The intersections between the solid black and gray lines ("not true" vs. "somewhat true") and between the gray and dashed black lines ("somewhat true" vs. "often true") are the thresholds that were calculated by adding the step location estimates to the overall item locations. In my example, Winsteps estimated the step parameters to be −1.41 logits and 1.41 logits. For the first BPI item the overall estimated item location was −1.83; thus, the first threshold was −1.83 + −1.41 = −3.24 and the second threshold was −1.83 + 1.41 = −0.42.

<Figure 3 about here>

Item–person map. Because the Rating Scale Model estimated one set of step locations across all of the items, the category probability curves for other items had the same shape as shown in Figure 3, but shifted to the left or to the right, depending on the overall item location.

The locations of all of the items in an item–person map are summarized in Figure 4. This map is similar to the one in Figure 2, but it now positions the items at the thresholds between the first and second response categories (the lower set of items in the map, whose labels end in .2) and between the second and third response categories (the higher set of items in the map, whose labels end in .3). The positioning of the items is not precise in the Winsteps display (the item thresholds and person locations can be exported for other graphics software), and the minimum logit value differs between Figure 2 and Figure 4. Still, comparing the two displays offers general insights into the advantage of the expanded three-category over the collapsed

dichotomous coding. For instance, it is clear that the items in Figure 4 extend from a minimum logit of below −3 to a maximum logit of above 3, whereas in Figure 2 they extend from just below −2 to 2. The thresholds in Figure 4 also overlap in the mid-range (between about −0.5 and 0.5 logits), which helps cover gaps (i.e., the thresholds between the first and second categories of the hardest items [Items 14, 16, 9, 15, and 5] overlapped the thresholds between the second and third categories of the easiest items [Items 4, 1, 8, 6, and 12]).

<Figure 4 about here>

Information functions. Item and test information functions help quantify the benefits of the expanded response structure. In CTT a single estimate of measurement error and of reliability applies to the entire instrument. In contrast, in IRT the error of each item's and each person's location is estimated. The inverse of this error is referred to as *information*. In IRT it is common to use item and test information functions to summarize the quality of items and of instruments. An item's information varies across the levels of the latent trait. In the Rasch model it is defined by multiplying the conditional probability of endorsing the item with the conditional probability of not endorsing the item, $p(1 - p)$, where p is calculated by substituting the estimated item location, $\hat{\beta}_i$, into Equation 1 and varying the level of the person location, θ_s, along the logit scale. For the Rasch model, item information reaches its maximum when the probability is .5 and information is $.5 \times (1 - .5) = .5 \times .5 = .25$. At other probabilities, the value will be smaller (e.g., $.75 \times .25 = .19$ and $.90 \times .10 = .09$) to a minimum of 0. Because the probability of .5 occurs when an item is at a person's location (i.e., when $\theta_s - \beta_i = 0$ in Equation 1), each item will be maximally informative for persons located at the item's difficulty level. As the distance between a person and the item increases, the item is less informative. Intuitively, for instance, a math item that is very easy or very hard tells us less about a person's math ability than would an item that is

just at the level of the person's current math competency. The formula for item information in the Rating Scale Model is more complex (de Ayala, 2009, p. 200), and the shape of the information function for this model depends on the number and spacing of thresholds, although, in general, increasing the number of categories will increase information (intuitively, the thresholds allow the item to cover a broader range of the latent trait).

Panel A of Figure 5 shows example item information functions for the dichotomous and three-category BPI items from Winsteps. For other items, the functions follow the displayed shapes but shift to the left or right depending on the item's difficulty level. The item information function peaks at the item's difficulty level for the dichotomous Rasch model; for the Rating Scale Model it peaks near the item's difficulty level (de Ayala, 2009, p. 201; Dodd & Koch, 1987). The information then fell off (e.g., for dichotomously scored items, when a person is located two logits from an item, information has fallen by more than half). In Panel A of Figure 5 it is clear that the three-category response structure (solid black line) offers more information than the dichotomous coding (dashed black line), having a higher and broader peak. During measure development, such information functions could be used to plan and evaluate the number of response categories. For instance, assuming items were well placed, fewer three-category items than dichotomous items would be needed to cover the full range of a latent construct (and, likewise, well-defined four-category items could cover more ground than three-category items). During initial studies, categories could be evaluated with IRT models to ensure they are achieving planned goals. For instance, respondents may not use all categories as expected, and in such cases they may need to be relabeled, reduced, or expanded. Additional Winsteps results, and advanced models, can help identify and remedy such situations (e.g., in one study, IRT models revealed that respondents did not seem able to distinguish between the categories of

"rarely" and "sometimes" on a 5-point scale of children's behaviors; Gordon et al., 2014).

<Figure 5 about here>

Test information was calculated by summing across item information at each point along the latent trait. Panel B of Figure 5 shows the test information as calculated by Winsteps, recentered to reflect the items' position relative to the estimated mean of the sample children (with items centered at 1.58 logits for the dichotomous recoding and 2.98 for the three-category response structure). It is not surprising that the graph shows that the BPI test as a whole provided more information based on the three-category (solid black line) than the dichotomous coding (dashed black line). The test information curve also demonstrates how quickly the information fell off from the maximum. By 2.5 logits out from the center, information on the dichotomously scored test had fallen by half. For the multicategory test, information had dropped in half by about 3 logits out from the center.

2PL model. In this final section I end my illustration by returning to the 2PL model and showing the implications of allowing the discrimination parameter to vary in the dichotomous scoring. To help solidify understanding, I show the item parameter estimates as well as example ICCs. I also present test information, comparing information when I allow the discriminations to vary and when I constrain them to be equal.

Parameter estimates and ICCs. To begin, in Figure 6, Panel A, I graphed the item difficulties that were shown in Table 2 for the 1PL and 2PL *gsem* models (using the 2PL parameterization form Column 4 of the table, with the item discriminations constrained to average 1). The graph shows that these 1PL and 2PL item difficulties are ordered similarly, although there was some reordering within sets of items. For instance, Items 3, 11, and 13 are positioned similarly relatively to the other items but differently in relation to one another (e.g.,

Item 13 was the hardest of the three in the 1PL model, but Item 11 was the hardest of the three in the 2PL model). Because the 2PL model allowed the discrimination parameters to vary, looking beyond just the inflection points was also important. Panels B and C of Figures 6 depict the full ICCs under the 1PL and 2PL models for two items whose discriminations differed from 1: (a) Item 1, whose discrimination parameter was 0.82, and (b) Item 2, whose discrimination parameter was 0.65. The graphs help reveal that the implications of allowing different discriminations is minimal in Panel B but more substantial in Panel C.

<Figure 6 about here>

Panel D of Figure 6 further shows the implications of the varying discriminations for the relative difficulty of items. Here I show ICCs for two different items under the 2PL model, one with the lowest discrimination value (Item 2, value of 0.65) and the other with the highest discrimination value (Item 7, value of 1.25). The steeper slope for Item 7 (dashed black line) versus Item 2 (solid black line) is readily apparent. The probability of endorsing Item 7 increased slowly at first, and then rose sharply starting at about 0 logits. In contrast, the probability of endorsing Item 2 increased steadily throughout, appearing nearly linear in much of the region graphed. The two items had similar difficulties (inflection points), values of 2.11 and 2.25 from Column 4 in Table 2. Like in the Rasch model, these points still reflect the location on the latent trait where a mother had a 50–50 chance of endorsing the dichotomously scored items, but they no longer reflect the relative order of the items across the full latent continuum. In other words, Panel D shows that up to about 1.0 logits mothers were more likely to report that the child was high strung than didn't get along with other children (the solid black line for Item 2 was higher than the dashed black line for Item 7); this order is reversed on the right hand side of the panel. Some analysts might prefer the 2PL model allowing such different discriminations when

estimating children's location on the latent trait. Other analysts might want to revisit the concept and items to try to explain this reversal and potentially modify the item to avoid the issue in future instrument development.

Item and test information function. In the 1PL and 2PL models the item information is calculated similarly as in the Rasch model, but the product $p(1-p)$ is multiplied by the square of the item's discrimination, that is, $\alpha_i^2 p(1-p)$, where α_i is estimated on the basis of Equation 2. For the 1PL model, the multiplier is the same for all items, given the discriminations are constrained to be equal. Equating further puts information on equal footing (e.g., de Ayala, 2009, p. 122).

Figure 7 shows the 2PL item information functions for Items 2 and 7 (top panel) and the test information functions for the 1PL and 2PL models (bottom panel, with both models parameterized by constraining the latent variance to 1). Figure 7, Panel A, makes clear the higher peak of information for the item with the steeper slope (solid black curve) but focused in a narrower range of the latent trait. In contrast, the item with the less steep slope has a lower peak in information (dashed curve) but is more equally informative across the full range of the latent trait. In fact, Item 2 offers somewhat more information than Item 7 in the extremes. IRT analysts sometimes convert test information to more familiar reliability units by the equation $I/(I+1)$, where I is the information value. A horizontal line at Point 4 in Figure 7, Panel B, reflects the point at which reliability equals .80, because $4/(4+1) = 4/5 = .80$. This line shows that both the 1PL and 2PL models offer above .80 reliability in a limited range, from logit values of about -0.5 to 2.5, reflecting the fact that the items were relatively "hard" for this sample (there were few items in the low to moderate range of behavior problems).

<Figure 7 about here>

Technical Details and Model Extensions

As a basic introduction, I focused this article on three models—the Rasch, Rating Scale, and 2PL models—and a subset of key concepts and techniques. In this final section I discuss some technical details and extensions that readers may encounter in the literature or during their own IRT analyses. Other details and extensions are available in the references I cite.

Dimensionality and local dependence revisited. Advanced multidimensional Rasch and IRT models are increasingly available (see Chapter 12 of Bond & Fox, 2007, for an accessible discussion, and Gibbons et al., 2007, and Muraki & Carlson, 1995, for more detailed treatments; see also Wu, Adams, Wilson, & Haldane, 2007, for flexible software). Factor-analytic models have also been extended to dichotomous and polytomous outcomes (Bock et al., 1988; B. Muthén, 1983; Skrondal & Rabe-Hesketh, 2007; Wirth & Edwards, 2007), and the generalized structural equation modeling approach could incorporate additional dimensions. More sophisticated models are also available to capture what have been called "testlets" of dependent items (Steinberg & Thissen, 1996), to model multidimensionality (Wang et al., 2005), and to model multilevel dependencies (Jiao, Kamata, Wang & Jin, 2012),

Item linking and differential item functioning. IRT models are well suited to help scholars test whether items function differently across subgroups, contexts, time, or instrument forms (Osterlind & Everson, 2009; Thissen, Steinberg, & Wainer, 1993). When some common items exist and operate similarly across groups/replicates, then IRT models can be used to link scores. Doing so can produce appropriately equated scales for analyzing change over time and for conducting integrative or coordinated data analyses (Bauer & Hussong, 2009; Curran et al., 2008; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009). As needed, IRT models can allow parameters to differ for some items across these groups/replicates and adjust scores

appropriately. The differently functioning items can also be scrutinized to help scholars understand why and how measures vary across subpopulations or with development (Hitchcock et al., 2005; Hui & Triandis, 1985).

Sample sizes and power. As with most statistical procedures, researchers will want to know the answer to the question "How many cases and how many items do I need for IRT modeling?" As is typical, the best response is "It depends." Particularly important are the assumptions of maximum-likelihood estimation, limits of small cell sizes, and strategies for thinking about statistical power. Given that IRT methods draw on various maximum-likelihood estimation techniques, they rely on asymptotic properties that hold as sample sizes approach infinity. Researchers are often advised to use samples with at least 200 cases for maximum-likelihood techniques, more for complex models or skewed distributions (Long, 1997). Among items that are at the extremes of the latent distribution (very rare or very common in the full population) or are not well targeted at a particular sample (difficult or easy for a particular set of respondents), no or few cases may fall in some cells unless the sample size is large. At the extreme, parameters cannot be estimated for items that have no variation in the sample, or for persons who have no variation across the items.

Statistical power is also a consideration, although not all of the tools discussed above rely on hypothesis testing. The classic approach used by Rasch purists is quite interpretive, with analysts using multiple pieces of information to make decisions about the evidence regarding whether the data as a whole and certain items and persons are consistent with the model. Especially in the instrument development stage Rasch analysts may follow an iterative process in which items, or persons, are omitted one at a time with a goal toward identifying the set that work best together to define the dimension. Often, additional information, such as scrutiny of

item content by substantive experts and focus groups or follow-up interviews with respondents, is used in this process. This iterative process requires an item pool that is large enough that items can be removed and replaced, as needed. The Rasch community has developed conventions about sample sizes that may be helpful to analysts (Linacre, 1994), and some strategies have been published on power specifically for Rasch models (Draxler, 2010; Guilleux, Blanchin, Hardouin, & Sebille, 2014) and more broadly for using Monte Carlo techniques to estimate power for complex models (L. K. Muthén & Muthén, 2002).

Complex sampling designs. Perhaps in part because early Rasch models emphasized the separation of item and person statistics—that the relative location of items did not depend on the sample—there is limited discussion of complex sampling designs in the Rasch literature. Built-in options to adjust for complex sampling designs have also not historically been standard across latent variable—structural equation modeling—or IRT software, although they have been growing (Asparhouhov, 2005; Stapleton, 2006). Winsteps can incorporate person-weights, which could include sampling weights from complex sampling designs. The weights are interpreted like frequency weights, so researchers should rescale the weights to represent the total sample size (e.g., divide by the average weight; Gordon, 2012, pp. 123–127). The Stata *gsem* command can incorporate design features into the model (e.g., by treating primary sampling units as a level and using covariates for oversampled and stratified units). Neither package fully incorporates design-based approaches to complex sampling designs, however (including adjustments to standard errors for primary sampling units and strata).

Discussion

IRT offers family scientists many tools to add to their arsenal when examining the precision and validity of an instrument. Although some of the insights gleaned through IRT might also be

gained with CTT approaches, the orientation of IRT can help researchers think explicitly about the ordering of items along the underlying dimension and the reliability of measurement for persons with varying levels of the underlying construct. By taking this perspective, scholars may be able to write better items when developing a new measure or refining an existing measure, including by more precisely understanding the implications of the match between items and characteristics of the population that will use the scale (e.g., implications for person location estimates and for item and test information of clinical samples with scores focused in a particular range of the underlying construct or representative samples with scores spread across the continuum). When this process is embedded in a complete conceptual framework that defines the underlying construct in detail and makes a conceptual case for various items' positions in relation to that definition, resulting studies of scale development can elevate to the level of substantive studies of associations among constructs.

How can family scientists gain the skills needed in order to leverage these approaches in their own work? An increasing number of textbooks cover IRT (Bond & Fox, 2007; de Ayala, 2009; Embretson & Reise, 2000; Reise et al., 2005; Wilson, 2005). Summer courses on IRT are also offered at places such as the Summer Program in Quantitative Methods of Social Research at the University of Michigan's Interuniversity Consortium for Political and Social Research or the Summer StatsCamp sponsored by the Institute for Measurement, Methodology, Analysis and Policy at Texas Tech University. Online courses are also available, such as the University of Illinois at Chicago's course in Item Response Theory and Rasch Measurement offered through its online Measurement, Evaluation, Statistics, and Assessment master's program. Such courses can train scholars in Winsteps and other software for IRT (e.g., Cai, 2013; Cai, Du Toit, & Thissen, 2011; Paek & Han, 2012) as well as in strategies for estimation within the popular

Mplus program (L. K. Muthén & Muthén, 2013) or through embedding basic IRT models in multilevel models (Raudenbush et al., 2003).

In short, IRT has the potential to help family scientists gain precision and validity in the measures used to test their theories, thereby reducing fuzzy boundaries among constructs and increasing power to detect associations. Publications that conceptualize and test new measures, refine existing measures, or replicate results in new contexts recognize the central role that such work has within the broader social science enterprise.

REFERENCES

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and Profile*. Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington: University of Vermont, Research Center for Children, Youth, & Families.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology. *American Psychologist, 63,* 32–50. doi:10.1037/0003-066X.63.1.32

Aiken, L. S., West, S. G., Sechrest, L. B., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology. *American Psychologist, 45,* 721–734. doi:10.1037/0003-066X.45.6.721

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42,* I7–I16. doi:10.1097/01.mlr.0000103528.48582.7c

Asparhouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12,* 411–434. doi:10.1207/s15328007sem1203_4

Barnes, G. M., Hoffman, J. H., Welte, J. W., Farrell, M. P., & Dintcheff, B. A. (2006). Effects of parental monitoring and peer deviance on substance use and delinquency. *Journal of Marriage and Family, 68,* 1084–1104. doi:10.1111/j.1741-3737.2006.00315.x

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14,* 101–125. doi:10.1037/a0015583

Bingenheimer, J. B., Raudenbush, S. W., Leventhal, T., & Brooks-Gunn, J. (2005). Measurement equivalence and differential item functioning in family psychology. *Journal of Family Psychology, 19,* 441–455. doi:10.1037/0893-3200.19.3.441

Black, R. A., & Butler, S. F. (2012). Using the GLIMMIX procedure in SAS 9.3 to fit a standard dichotomous Rasch and hierarchical 1-PL IRT model. *Applied Psychological Measurement, 36,* 237–248. doi:10.1177/0146621612441857

Blinkhorn, S. F. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50,* 175–185. doi:10.1111/j.2044-8317.1997.tb01139.x

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280. doi:10.1177/014662168801200305

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Bridges, M., Cohen, S. R., McGuire, L. W., Yamada, H., Fuller, B., Mireles, L., & Scott, L. (2012). "Bien educado": Measuring the social behaviors of Mexican American children. *Early Childhood Research Quarterly, 27,* 555–567. doi:10.1016/j.ecresq.2012.01.005

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Browning, C. R. (2002). The span of collective efficacy: Extending social disorganization theory to partner violence. *Journal of Marriage and Family, 64,* 833–850. doi:10.1111/j.1741-3737.2002.00833.x

Browning, C. R., & Burrington, L. A. (2006). Racial differences in sexual and fertility attitudes in an urban setting. *Journal of Marriage and Family, 68,* 236–351. doi:10.1111/j.1741-3737.2006.00244.x

Buehler, C. (2006). Parents and peers in relation to early adolescent problem behavior. *Journal of Marriage and Family, 68,* 109–124. doi:10.1111/j.1741-3737.2006.00237.x

Cai, L. (2013). flexMIRT Version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago: Scientific Software International.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6). Retrieved from www.jstatsoft.org/v48/i06/paper

Chen, W., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265–289. doi:10.2307/1165285

Cho, Y. I., Martin, M. J., Conger, R. D., & Widaman, K. F. (2010). Differential item functioning on antisocial behavior scale items for adolescents and young adults from single-parent and two-parent families. *Journal of Psychopathology and Behavioral Assessment, 32,* 157–168. doi:10.1007/s10862-009-9145-1

Chorpita, B. F., Reise, S., Weisz, J. R., Grubbs, K., Becker, K. D., Krull, J. L., & The Research Network on Youth Mental Health. (2010). Evaluation of the Brief Problem Checklist: Child and caregiver interviews to measure clinical progress. *Journal of Consulting and Clinical Psychology, 78,* 526–536. doi:10.1037/a0019602

Cole, J. C., Rabin, A. S., Smith, T. L., & Kaufman, A. S. (2004). Development and validation of a Rasch-derived CES–D short form. *Psychological Assessment, 16,* 360–372. doi:10.1037/1040-3590.16.4.360

Coley, R. L., Ribar, D., & Votruba-Drzal, E. (2011). Do children's behavior problems limit poor women's labor market success? *Journal of Marriage and Family, 73,* 33–45. doi:10.1111/j.1741-3737.2010.00787.x

Conrad, K. J., Conrad, K. M., Mazza, J., Riley, B. B., Funk, R., Stein, M. A., & Dennis, M. L. (2012). Dimensionality, hierarchical structure, age generalizability, and criterion validity of the GAIN's behavioral complexity scale. *Psychological Assessment, 24,* 913–924. doi:10.1037/a0028196

Conrad, K. J., Riley, B. B., Conrad, K. M., Chan, Y., & Dennis, M. L. (2010). Validation of the Crime and Violence Scale (CVS) against the Rasch measurement model including differences by gender, race, and age. *Evaluation Review, 34,* 83–115. doi:10.1177/0193841X10362162

Culpepper, S. A. (2013). Reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distributions. *Applied Psychological Measurement, 37,* 201–225. doi:10.1177/0146621612470210

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology, 44,* 365–380. doi:10.1037/0012-1649.44.2.365

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

DeWalt, D. A., Thissen, D., Stucky, B. D., Langer, M. M., DeWitt, E. M., Irwin, D. E., . . . Varni, J. W. (2013). PROMIS Pediatric Peer Relationships Scale: Development of a peer relationships item bank as part of social health measurement. *Health Psychology, 32,* 1093–1103. doi:10.1037/a0032670

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement, 11,* 371–384. doi:10.1177/014662168701100403

Draxler, C. (2010). Sample size determination for Rasch model tests. *Psychometrika, 75,* 708–724. doi:10.1007/s11336-010-9182-4

Elliott, D. S., Huizinga, D., & Ageton, S. S. (1985). *Explaining delinquency and drug use*. Beverly Hills, CA: Sage.

Embretson, S. E. (1986). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20,* 201–212. doi:10.1177/014662169602000302

Embretson, S. E., & Reise, S. R. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fincham, F. D., & Rogge, R. (2010). Understanding relationship quality: Theoretical challenges and new tools for assessment. *Journal of Family Theory and Review, 2,* 227–242. doi:10.1111/j.1756-2589.2010.00059.x

Fomby, P., & Bosick, S. J. (2013). Family instability and the transition to adulthood. *Journal of Marriage and Family, 75,* 1266–1287. doi:10.1111/jomf.12063

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78,* 350–365. doi:10.1037/0022-3514.78.2.350

Funk, J. L., & Rogge, R. D. (2007). Testing the ruler with item response theory: Increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index. *Journal of Family Psychology, 21,* 572–583. doi:10.1037/0893-3200.21.4.572

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31,* 4–19. doi:10.1177/0146621606289485

Gibson, C. L., Ward, J. T., Wright, J. P., Beaver, K. M., & Delisi, M. (2010). Where does gender fit in the measurement of self-control? *Criminal Justice and Behavior, 37,* 883–903. doi:10.1177/0093854810369082

Gordon, R. A. (2012). *Applied statistics for the social and health sciences*. New York: Routledge.

Gordon, R. A., Colwell, N., Fujimoto, K., Abner, K., Kaestner, R., Wakschlag, L., & Korenman, S. (2014). *Rating young children's behaviors: Similarities and differences among parents, family child care providers and teachers*. Chicago: Institute of Government and Public Affairs.

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS–R with implications for measure of child care quality and relations to child development. *Developmental Psychology, 49,* 146–160. doi:10.1037/a0027899

Guilleux, A., Blanchin, M., Hardouin, J., & Sebille, V. (2014). Power and sample size determination in the Rasch model. *PLOS One, 9,* e83652. doi:10.1371/journal.pone.0083652

Hitchcock, J. H., Nastasi, B. K., Dai, D. Y., Newman, J., Jayasena, A., Bernstein-Moore, R., . . . Varjas, K. (2005). Illustrating a mixed-method approach for validating culturally specific constructs. *Journal of School Psychology, 43,* 259–278. doi:10.1016/j.jsp.2005.04.007

Ho, A. (2014, March). Item response theory. Workshop conducted at the meeting of the Society for Research in Educational Effectiveness, Washington, DC.

Houts, C. R., & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement, 37,* 541–562. doi:10.1177/0146621613491456

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16,* 131–152. doi:10.1177/0022002185016002001

IBM. (2014). *Item response theory/Rasch models in SPSS statistics*. Retrieved from www-01.ibm.com/support/docview.wss?uid=swg21488442

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66,* 109–132. doi:10.1007/BF02295736

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49,* 82–100. doi:10.1111/j.1745-3984.2011.00161.x

Johnson, S., Li, J., Kendall, G., Strazdins, L., & Jacoby, P. (2013). Mothers' and fathers' work hours, child gender, and behavior in middle childhood. *Journal of Marriage and Family, 75,* 56–74. doi:10.1111/j.1741-3737.2012.01030.x

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36,* 347–387.

Kang, S., & Waller, N. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement, 29,* 87–105. doi:10.1177/0146621604272737

Krishnakumar, A., Buehler, C., & Barber, B. K. (2004). Cross-ethnic equivalence of socialization measures in European American and African American youth. *Journal of Marriage and Family, 66,* 809–820. doi:10.1111/j.0022-2445.2004.00054.x

Lambert, M. C., Schmitt, N., Samms-Vaughan, M. E., An, J. S., Fairclough, M., & Nutter, C. A. (2003). Is it prudent to administer all items for each child behavior checklist cross-informant syndrome? Evaluating the psychometric properties of the Youth Self-Reported Dimensions with confirmatory factor analysis and item response theory. *Psychological Assessment, 15,* 550–568. doi:10.1037/1040-3590.15.4.550

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Linacre, J. M. (2012). *Winsteps® Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

McArdle, J. J., Grimm, K., Hamagami, F., Bowles, R., & Meredith, W. (2009). Modeling life-space growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14,* 126–149. doi:10.1037/a0015857

McLoyd, V. C., & Smith, J. (2002). Physical discipline and behavior problems in African American, European American, and Hispanic children: Emotional support as a moderator. *Journal of Marriage and Family, 64,* 40–53. doi:10.1111/j.1741-3737.2002.00040.x

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379–416. doi:10.1111/j.1745-3984.1996.tb00498.x

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73–90. doi:10.1177/014662169501900109

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22,* 43–65. doi:10.1016/0304-4076(83)90093-3

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599–620. doi:10.1207/S15328007SEM0904_8

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.) Los Angeles: Muthén & Muthén.

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Routledge.

Osborne, C., & McLanahan, S. (2007). Partnership instability and child well-being. *Journal of Marriage and Family, 69,* 1065–1083. doi:10.1111/j.1741-3737.2007.00431.x

Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance: I. Item response theory scaling. *Journal of Quantitative Criminology, 18,* 267–296. doi:10.1023/A:1016008004010

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks: Sage. doi:10.4135/9781412993913

Paek, I., & Han, K. T. (2012). IRTRO 2.1 for Windows. *Applied Psychological Measurement, 37,* 242–252. doi:10.1177/0146621612468223

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent–child relationships, and behavior problems in children. *Journal of Marriage and the Family, 48,* 295–307. doi:10.2307/352397

Peterson, R. A. (2000). A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing Letters, 11,* 261–275. doi:10.1023/A:1008191211004

Piquero, A. R., Macintosh, R., & Hickman, M. (2002). The validity of a self-reported delinquency scale: Comparisons across gender, age, race and place of residence. *Sociological Methods and Research, 30,* 492–529. doi:10.1177/0049124102030004002

Raîche, G. (2005). Critical eigenvalue sizes (variances) in standardized residual principal components analysis. *Rasch Measurement Transactions, 19,* 1011–1012.

Rapport, M. D., LaFond, S. V., & Sivo, S. A. (2009). Unidimensionality and developmental trajectory of aggressive behavior in clinically-referred boys: A Rasch analysis. *Journal of Psychopathology and Behavioral Assessment, 31,* 309–319. doi:10.1007/s10862-008-9125-x

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A multivariate, multilevel Rasch model with application to self-reported criminal behavior. *Sociological Methodology, 33,* 169–211. doi:10.1111/j.0081-1750.2003.t01-1-00130.x

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications and promise in psychological research. *Current Directions in Psychological Science, 14,* 95–101. doi:10.1111/j.0963-7214.2005.00342.x

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566. doi:10.1037/0033-2909.114.3.552

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17,* 1–25.

Rocque, M., Posick, C., & Zimmerman, G. M. (2013). Measuring up: Assessing the measurement properties of two self-control scales. *Deviant Behavior, 34,* 534–556. doi:10.1080/01639625.2012.748619

SAS Institute. (2013). Proc IRT (Experimental). In *SAS/STAT 13.1 user's guide*. Cary, NC: Author.

Schwab, D. P. (1980). Construct validity in organizational behavior. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (pp. 2–43). Greenwich: JAI Press.

Shafer, A. B. (2006). Meta-analyses of the factor structures of four depression questionnaires: Beck, CES–D, Hamilton, and Zung. *Journal of Clinical Psychology, 62,* 123–146. doi:10.1002/jclp.20213

Sheu, C., Chen, C., Su, Y., & Wang, W. (2005). Using SAS PROC NLMIXED to fit item response theory models. *Behavior Research Methods, 37,* 202–218. doi:10.3758/BF03192688

Skrondal, A., & Rabe-Hesketh, S. (2007). Latent variable modeling: A survey. *Scandinavian Journal of Statistics, 34,* 712–745. doi:10.1111/j.1467-9469.2007.00573.x

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size consideration for polytomous data. *BMC Medical Research Methodology, 8,* 33. doi:10.1186/1471-2288-8-33

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3,* 25–40. doi:10.1080/10705519609540027

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 73–92). Maple Grove, MN: JAM Press.

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling, 13,* 28–58. doi:10.1207/s15328007sem1301_2

StataCorp. (2013). Stata: Release 13 [Computer software]. College Station, TX: Author.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1,* 81–97. doi:10.1037/1082-989X.1.1.81

Studts, C. R., & van Zyl, M. A. (2013). Identification of developmentally appropriate screening items for disruptive behavior problems in preschoolers. Journal of Abnormal Child Psychology, 41, 851–863). doi:10.1007/s10802-013-9738-8

Sweeten, G. (2012). Scaling criminal offending. *Journal of Quantitative Criminology, 28,* 533–557. doi:10.1007/s10940-011-9160-8

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408. doi:10.1007/BF02294363

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. In *Crime and justice 2000: Vol. 4. Measurement and analysis of crime and justice* (pp. 33–83). Washington, DC: U.S. Department of Justice.

Turney, K. (2011). Chronic and proximate depression among mothers: Implications for child well-being. *Journal of Marriage and Family, 73,* 149–163. doi:10.1111/j.1741-3737.2010.00795.x

Wang, W., Cheng, Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement, 65,* 5–27. doi:10.1177/0013164404268676

White, J. M., & Klein, D. M. (2008). *Family theories*. Thousand Oaks: Sage.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58–79. doi:10.1037/1082-989X.12.1.58

Wolfe, F. (2001). raschcvt [User-written Stata program]. Retrieved from www.winsteps.com/winman/datafromstatafiles.htm

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER Conquest, Version 2: Generalised items response modeling software [Computer software]. Victoria, Australia: ACER Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Table 1. *Item Fit Statistics From the Rasch Model as Estimated in Winsteps*

Item	Infit			Outfit	
	MNSQ	*zstd*		MNSQ	*zstd*
1. Has sudden changes in mood or feeling	1.11	4.6		1.10	1.6
2. Is rather high strung, tense, and nervous	1.17	5.8		1.26	4.4
3. Cheats or tells lies	1.05	2.0		1.04	0.8
4. Argues too much	0.96	−1.9		0.94	−1.1
5. Bullies or is cruel/mean to others	0.91	−2.4		0.78	−2.5
6. Is disobedient at home	0.95	−2.3		0.92	−2.0
7. Has trouble getting along with other children	0.90	−2.7		0.73	−3.3
8. Is impulsive or acts without thinking	0.95	−2.1		0.90	−2.4
9. Is not liked by other children	1.03	0.6		1.15	1.2
10. Has trouble getting mind off certain thoughts	1.15	5.2		1.23	3.8
11. Is restless, overly active, cannot sit still	1.10	3.9		1.20	4.3
12. Is stubborn, sullen, or irritable	0.89	−5.3		0.82	−4.7
13. Has strong temper and loses it easily	0.90	−4.1		0.83	−3.9
14. Breaks things deliberately	1.02	0.4		1.04	0.3
15. Is disobedient at school	0.94	−1.7		0.89	−1.3
16. Has trouble getting along with teachers	0.97	−0.6		0.82	−1.4

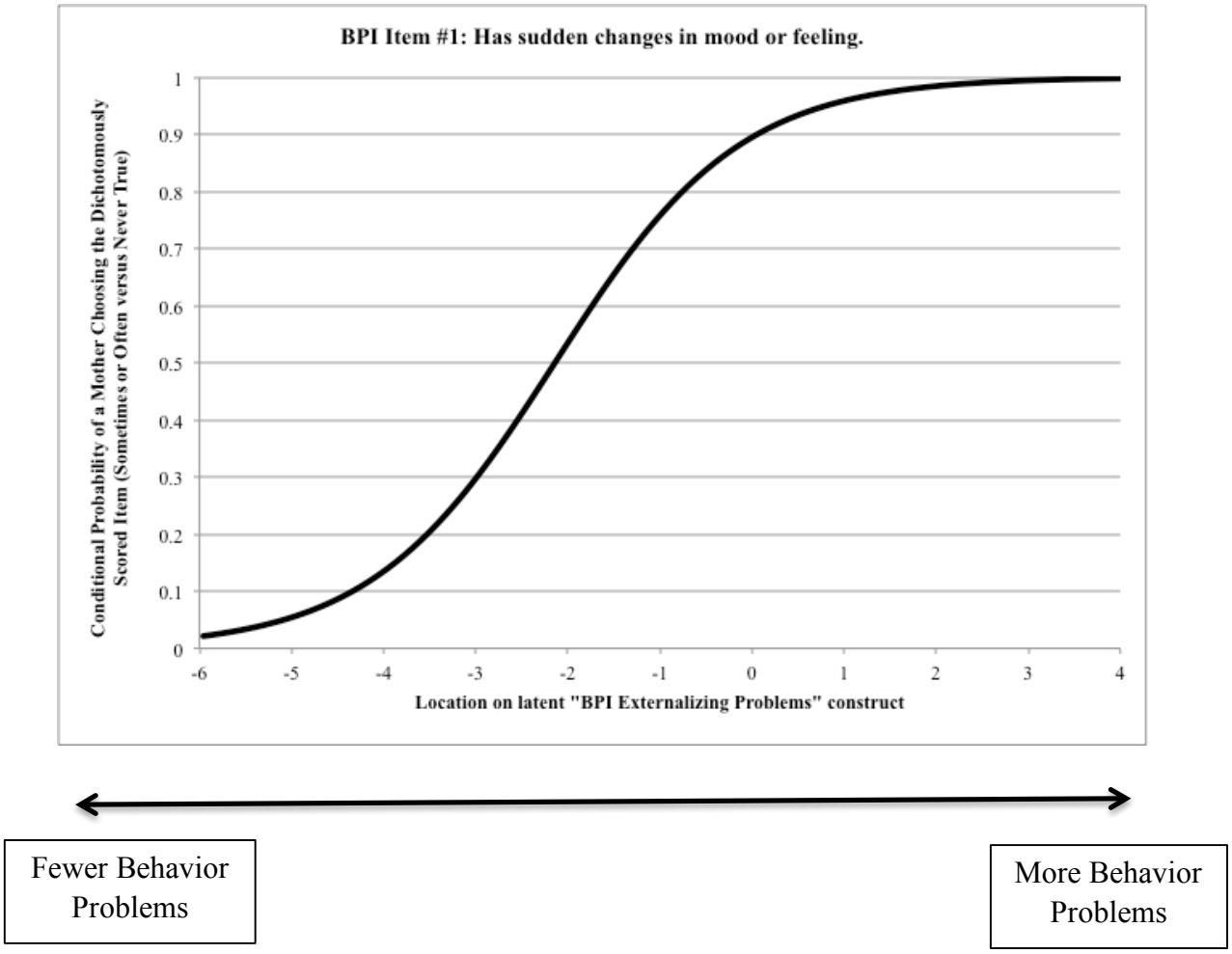
Note: n = 2,732. The model was estimated with Winsteps using the National Longitudinal Survey of Youth 1979 Behavior Problems Index externalizing items coded dichotomously (1 = *not true* and 2 = *sometimes true* or *often true*). *Infit* and *Outfit* are summary statistics based on the squared residuals. *Infit* is more sensitive to high residuals for persons located near the item's location. *Outfit* is more sensitive to high residuals for persons located far from the item's location. MNSQ = mean square (values expected to be between about 0.70 and 1.30); *zstd* = standardized value (values above 2 or 3 in absolute values are considered extreme, although sensitive to sample size; A. B. Smith et al., 2008).

Table 2. *Item Difficulty and Discrimination Estimates for Rasch/One-Parameter Logistic (1PL) and Two-Parameter-Logistic (2PL) Specifications*

Item	Rasch[a] (difficulty)	1PL[b] (difficulty)	2PL[c] (average of item discriminations = 1)		2PL[d] (variance of latent distribution = 1)	
			Discrimination	Difficulty	Discrimination	Difficulty
1. Has sudden changes in mood or feeling	−2.14	−0.45	0.82	−0.51	1.48	−0.28
2. Is rather high strung, tense, and nervous	0.13	1.63	0.65	2.11	1.18	1.17
3. Cheats or tells lies	−0.45	1.09	0.88	1.18	1.58	0.66
4. Argues too much	−2.05	−0.38	1.12	−0.36	2.01	−0.20
5. Bullies or is cruel/mean to others	1.11	2.55	1.24	2.37	2.24	1.31
6. Is disobedient at home	−1.20	0.39	1.08[e]	0.39	1.95	0.22
7. Has trouble getting along with other children	0.98	2.43	1.25	2.25	2.26	1.25
8. Is impulsive or acts without thinking	−1.03	0.54	1.02[e]	0.55	1.84	0.31
9. Is not liked by other children	1.63	3.03	0.93[e]	3.20	1.68	1.78
10. Has trouble getting mind off certain thoughts	0.11	1.61	0.68	2.02	1.23	1.12
11. Is restless, overly active, cannot sit still	−0.41	1.12	0.74	1.34	1.33	0.74
12. Is stubborn, sullen, or irritable	−1.11	0.47	1.24	0.44	2.23	0.24
13. Has strong temper and loses it easily	−0.40	1.13	1.18	1.07	2.13	0.60
14. Breaks things deliberately	2.10	3.47	0.96[e]	3.60	1.73	2.00
15. Is disobedient at school	0.94	2.39	1.10[e]	2.34	1.98	1.30
16. Has trouble getting along with teachers	1.78	3.17	1.09[e]	3.10	1.97	1.72
Average of item parameters	0.00	1.51	1.00	1.57	1.80	0.87
Person/latent mean	−1.58	0.00	0.00		0.00	
Person/latent variance	3.72	3.03	3.25		1.00	
Log likelihood		−19,627.82	−19,512.31		−19,512.31	
Akaike Information Criterion		39,289.64	39,088.62		39,088.62	
Bayesian Information Criterion		39,390.16	39,277.83		39,277.83	

Note: n = 2,732. Models are based on the National Longitudinal Survey of Youth 1979 Behavior Problems Index externalizing items coded dichotomously (*1 = not true* and *2 = sometimes true* or *often true*) estimated with Winsteps (first column) or Stata *gsem*. [a]Parameterized so item mean is 0 (as in Equation 1 in the article text, the discrimination values are implicitly 1). [b]Parameterized with latent mean constrained to 0, latent variance freely estimated, and all discrimination values equal 1. [c]Parameterized with latent mean constrained to 0, latent variance freely estimated, and discrimination values average 1. [d]Parameterized with latent mean constrained to 0, latent variance constrained to 1, and all discrimination values freely estimated. [e]Value does not differ significantly from 1 ($p > .05$).

FIGURE 1. ITEM CHARACTERISTIC CURVE BASED ON WINSTEPS OUTPUT: BEHAVIOR PROBLEMS INDEX (BPI) ITEM 1: "HAS SUDDEN CHANGES IN MOOD OR FEELING."



Note: The line reflects the probability of the mother endorsing the item (selecting that the characteristic is sometimes or often versus never true of the child, on the dichotomously recoded BPI item), conditional on the child's location of the latent externalizing problems construct. The latent construct is in logit units. The estimated item difficulty is −2.14 in Winsteps.

FIGURE 2. ITEM–PERSON MAP FROM WINSTEPS FOR BEHAVIOR PROBLEM INDEX ITEMS, DICHOTOMOUSLY RECODED.

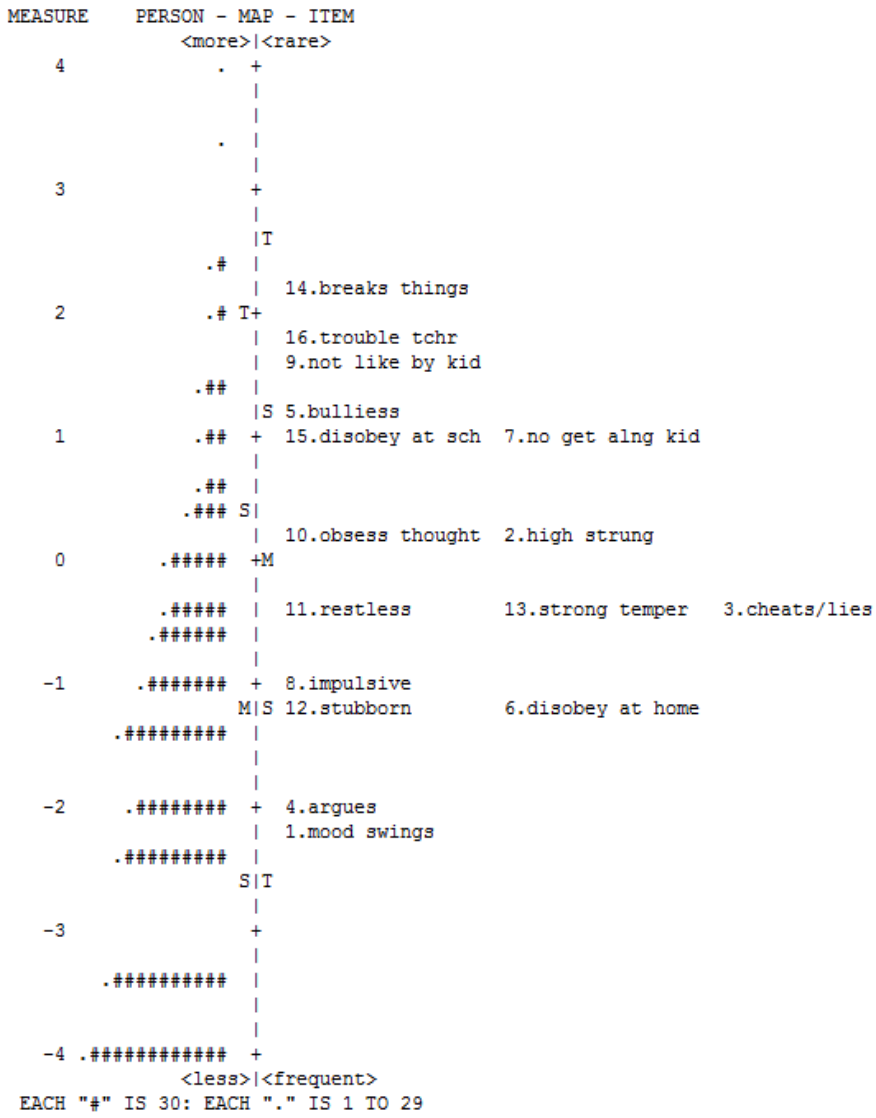
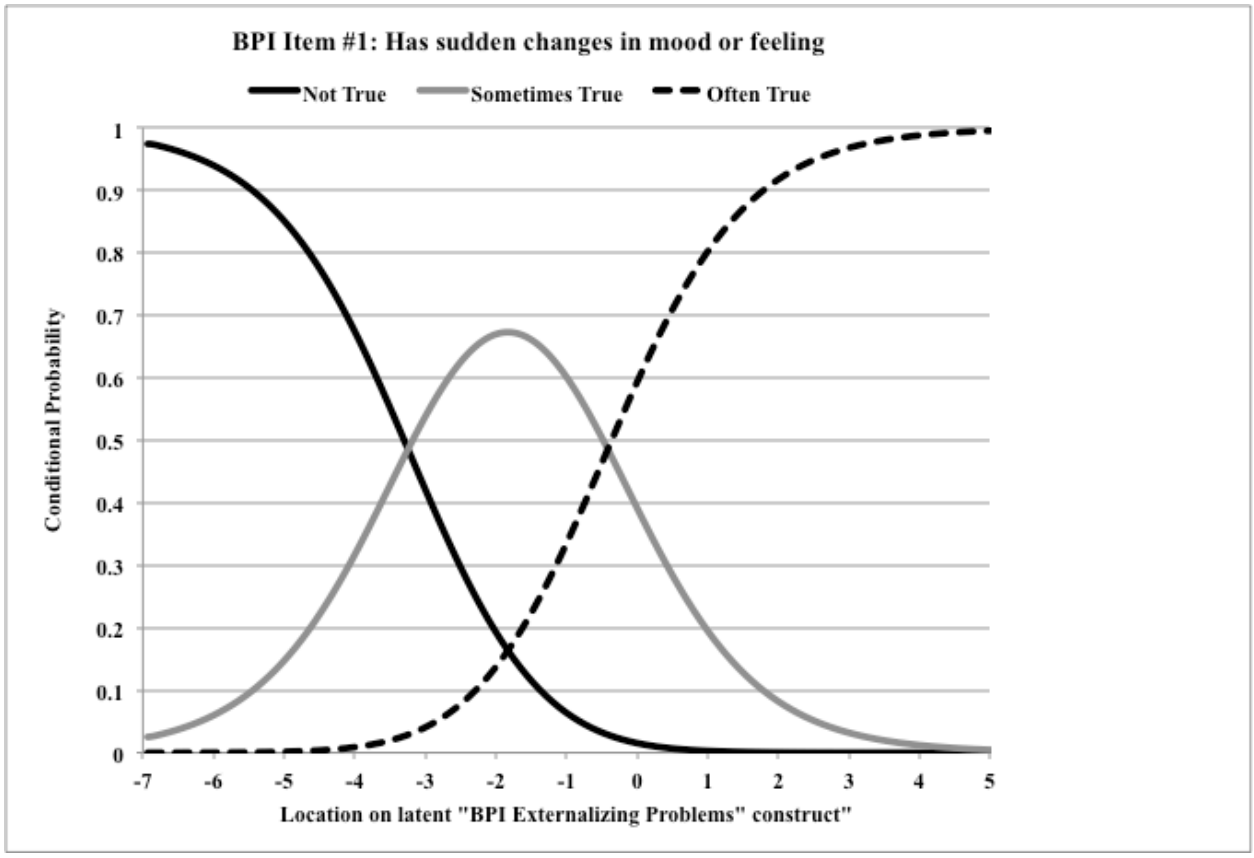


FIGURE 3. CATEGORY PROBABILITY CURVE BASED ON WINSTEPS OUTPUT: BEHAVIOR PROBLEMS INDEX (BPI) ITEM 1: "HAS SUDDEN CHANGES IN MOOD OR FEELING," WITH ORIGINAL THREE-CATEGORY RESPONSE OPTIONS.



Note: The solid black line, dark gray, and dashed black lines represent the probability of a response of *not true*, *sometimes true*, and *often true*, respectively, conditional on the child's location of the latent externalizing problems construct. The latent construct is in logit units. The overall item difficulty is −1.83, and the category steps are −1.41 and 1.41.

FIGURE 4. ITEM–PERSON MAP FROM WINSTEPS FOR BEHAVIOR PROBLEM INDEX ITEMS, WITH ORIGINAL THREE-CATEGORY RESPONSE OPTIONS.

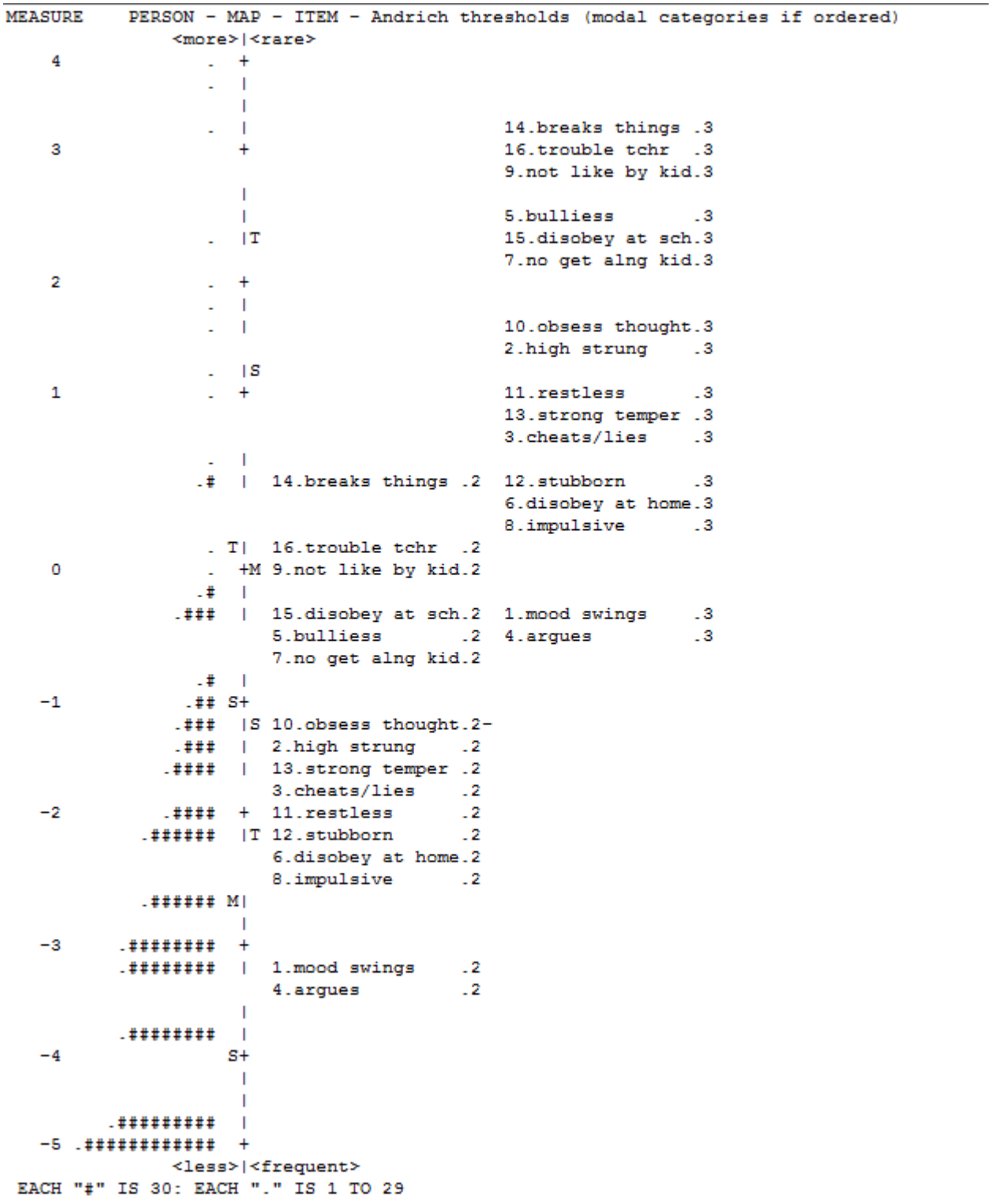
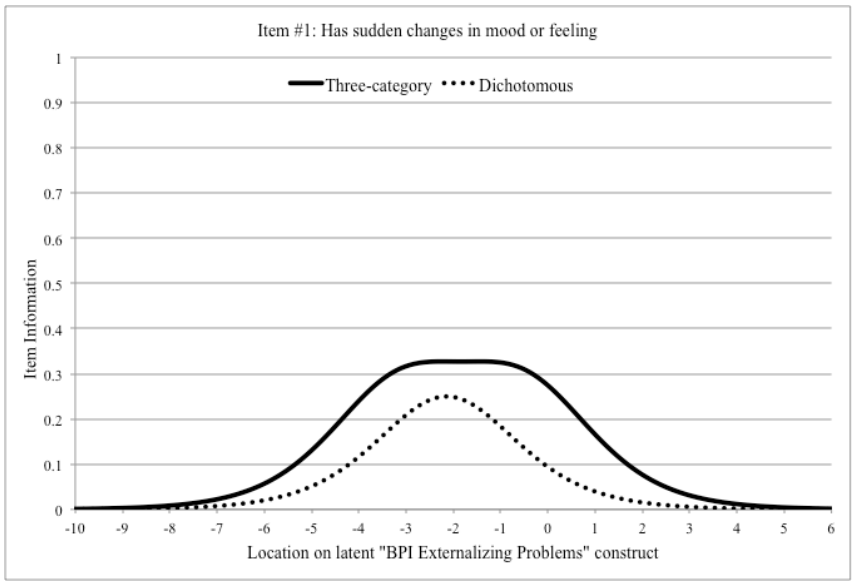
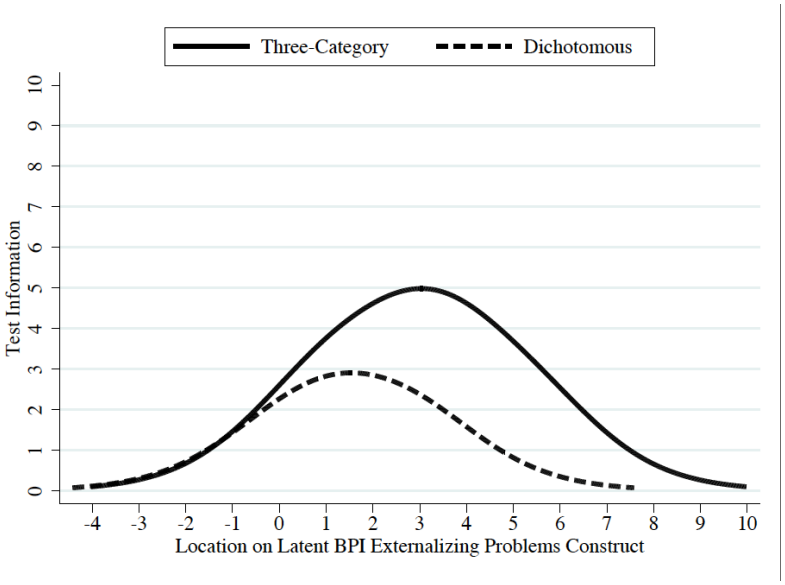


FIGURE 5. INFORMATION FUNCTIONS FOR ORIGINAL THREE-CATEGORY RESPONSE OPTIONS (RATING SCALE MODEL) AND DICHOTOMOUS RECODING (RASCH MODEL) BASED ON WINSTEPS RESULTS.

A. Example item information functions for Item 1



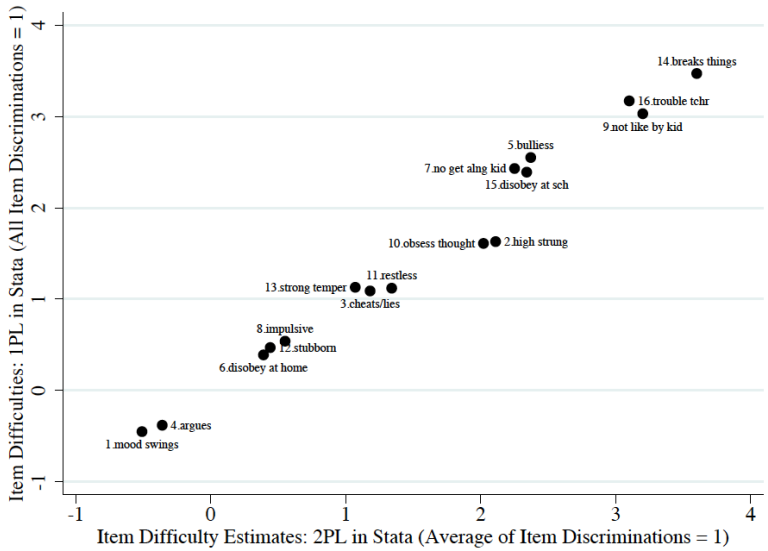
B. Test information functions for Rating Scale Model and Rasch Model



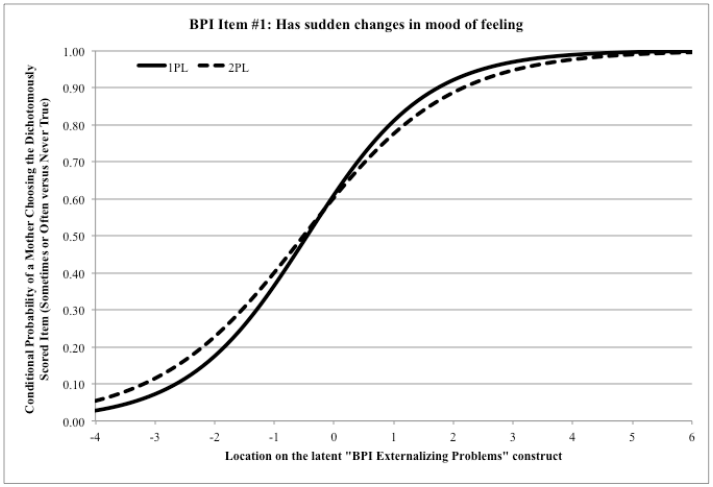
Note: In Panel A, the Winsteps item difficulty for Behavior Problems Index (BPI) Item 1: "Has sudden changes in mood or feeling," is −2.14 for the dichotomous coding, and the overall item location is −1.83 for the three-category response structure. In Panel B, the test information functions were recentered to reflect their position relative to the estimated mean of the sample children (with item means at 1.58 for the dichotomous recoding and at 2.98 for the three-category response structure).

FIGURE 6. COMPARISON OF ITEM DIFFICULTIES AND ITEM CHARACTERISTIC CURVES (ICCS) FOR THE RASCH AND TWO-PARAMETER LOGISTIC (2PL) MODEL FROM STATA'S *GSEM* COMMAND, USING THE DICHOTOMOUSLY RECODED ITEMS.

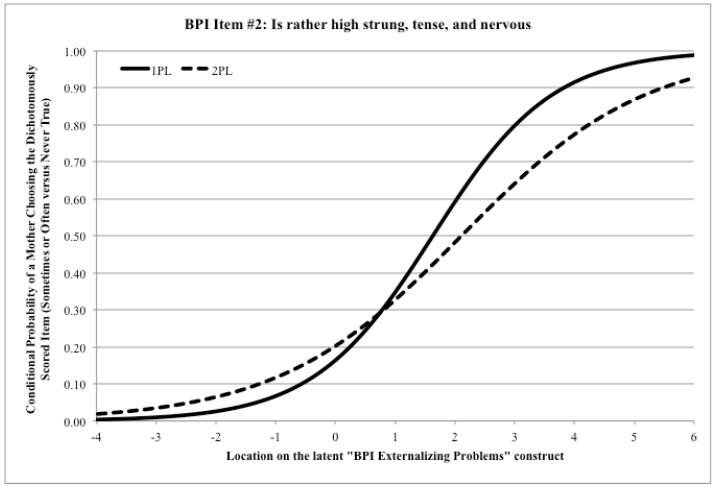
A. Graph of item difficulties in Rasch and 2PL model



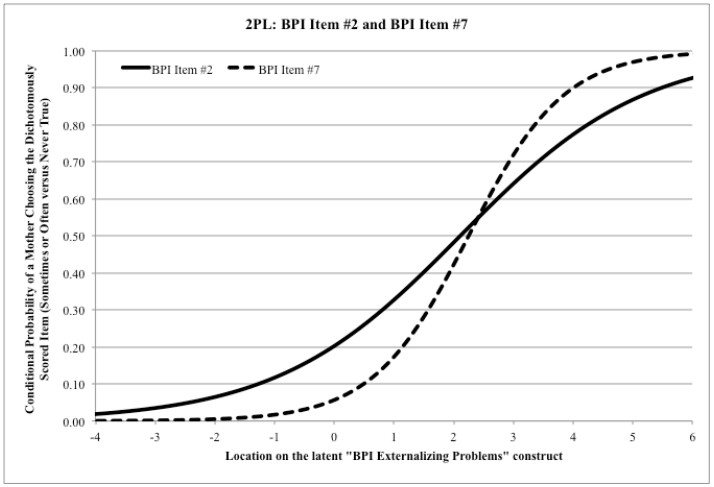
B. Example ICC in Rasch and 2PL model (Item 1)



C. Example ICC in Rasch and 2PL model (Item 2)



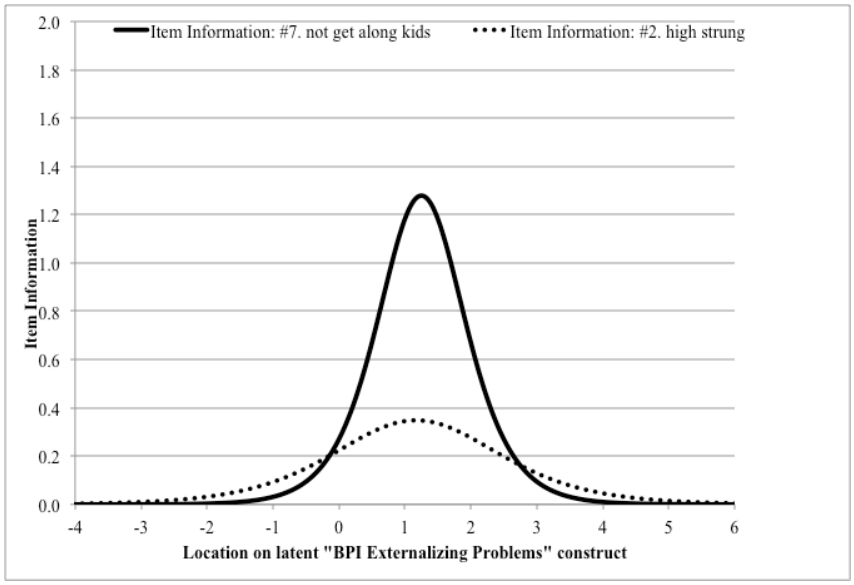
D. Example ICCs in 2PL model (Item 2 and Item 7)



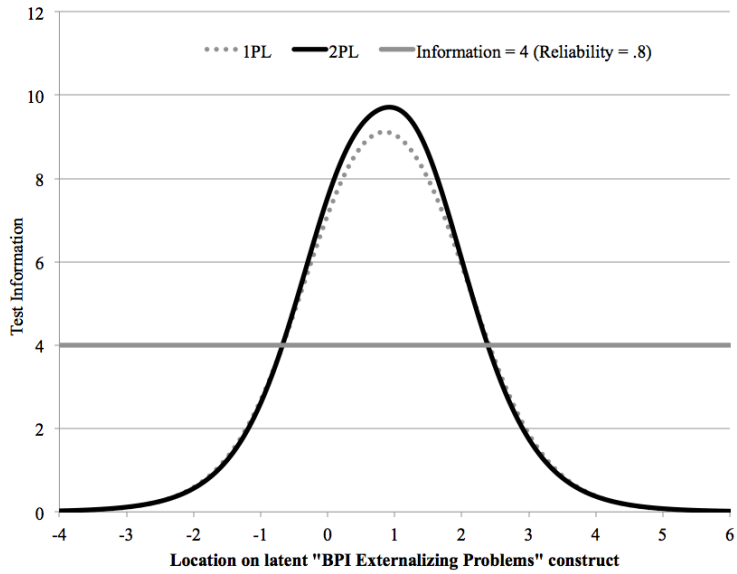
Note: 1PL = one-parameter logistic; BPI = Behavior Problems Index.

FIGURE 7. INFORMATION FUNCTIONS FOR THE TWO-PARAMETER LOGISTIC (2PL) MODEL BASED ON STATA'S *GSEM* COMMAND USING THE DICHOTOMOUSLY RECODED ITEMS.

A. Example item information functions



B. Test information functions



Note: In Panel A, the *gsem* item discrimination and difficulty are, respectively, 2.26 and 1.25 for Behavior Problems Index (BPI) Item 7 ("Has trouble getting along with other children") and 1.18 and 1.17 for BPI Item 2 ("Is high strung, tense, and nervous"). In Panel B, the test information functions are centered at the mean item difficulties of 0.87, relative to person-means of 0.

Appendix. *Item Wording and Descriptive Statistics for the Behavior Problems Index (BPI) Externalizing Items From the 2002 Survey of the Children of the National Longitudinal Study of Youth 1979 (NLSY79)*

	Response categories				Dichotomy
Item	Not true	Sometimes true	Often true		Sometimes or often true
1. Has sudden changes in mood or feeling	1,161 (43)	1,272 (47)	299 (11)		1,571 (58)
2. Is rather high strung, tense, and nervous	2,046 (75)	562 (21)	124 (5)		686 (25)
3. Cheats or tells lies	1,837 (67)	816 (30)	79 (3)		895 (33)
4. Argues too much	1,194 (44)	1,174 (43)	364 (13)		1,538 (56)
5. Bullies or is cruel/mean to others	2,326 (85)	372 (14)	34 (1)		406 (15)
6. Is disobedient at home	1,539 (56)	1,096 (40)	97 (4)		1,193 (44)
7. Has trouble getting along with other children	2,295 (84)	410 (15)	27 (1)		437 (16)
8. Is impulsive or acts without thinking	1,606 (59)	996 (36)	130 (5)		1,126 (41)
9. Is not liked by other children	2,435 (89)	272 (10)	25 (1)		297 (11)
10. Has trouble getting mind off certain thoughts	2,037 (75)	608 (22)	87 (3)		695 (25)
11. Is restless, overly active, cannot sit still	1,853 (68)	697 (26)	182 (7)		879 (32)
12. Is stubborn, sullen, or irritable	1,572 (58)	1,009 (37)	151 (6)		1,160 (42)
13. Has strong temper and loses it easily	1,854 (68)	727 (27)	151 (6)		878 (32)
14. Breaks things deliberately	2,512 (92)	202 (7)	18 (1)		220 (8)
15. Is disobedient at school	2,285 (84)	407 (15)	40 (1)		447 (16)
16. Has trouble getting along with teachers	2,461 (90)	243 (9)	28 (1)		271 (10)

Note: Data are unweighted sample sizes. Numbers in parentheses are percentages. Mothers were asked, "The following statements are about behavior problems many children have. For each item, think about [child's] behavior over the last three months. Then indicate whether the statement is often true, sometimes true, or not true." Items are renumbered for the pedagogical example and thus differ from the item numbers in the NLSY79 documentation. Results are based on 2,732 mothers who responded to all items. I excluded 135 mothers who did not respond to the final two items because children had not attended school. Other item-level missing data were sporadic and minimal (24 mothers, with one to eight mothers demonstrating unique missing data patterns).